ETH zürich

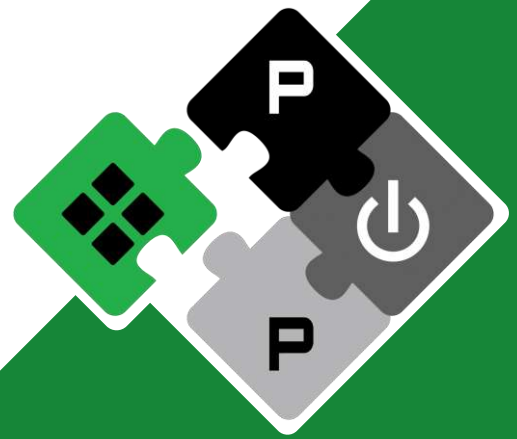
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Designing Digital Platforms for Wearable Contextual Intelligence

Luca Benini

lbenini@iis.ee.ethz.ch

luca.benini@unibo.it



@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

PULP Platform

Open Source Hardware, the way it should be!

Embodied (Physical) AI: Artificial Intelligence Everywhere

Nano-Drone

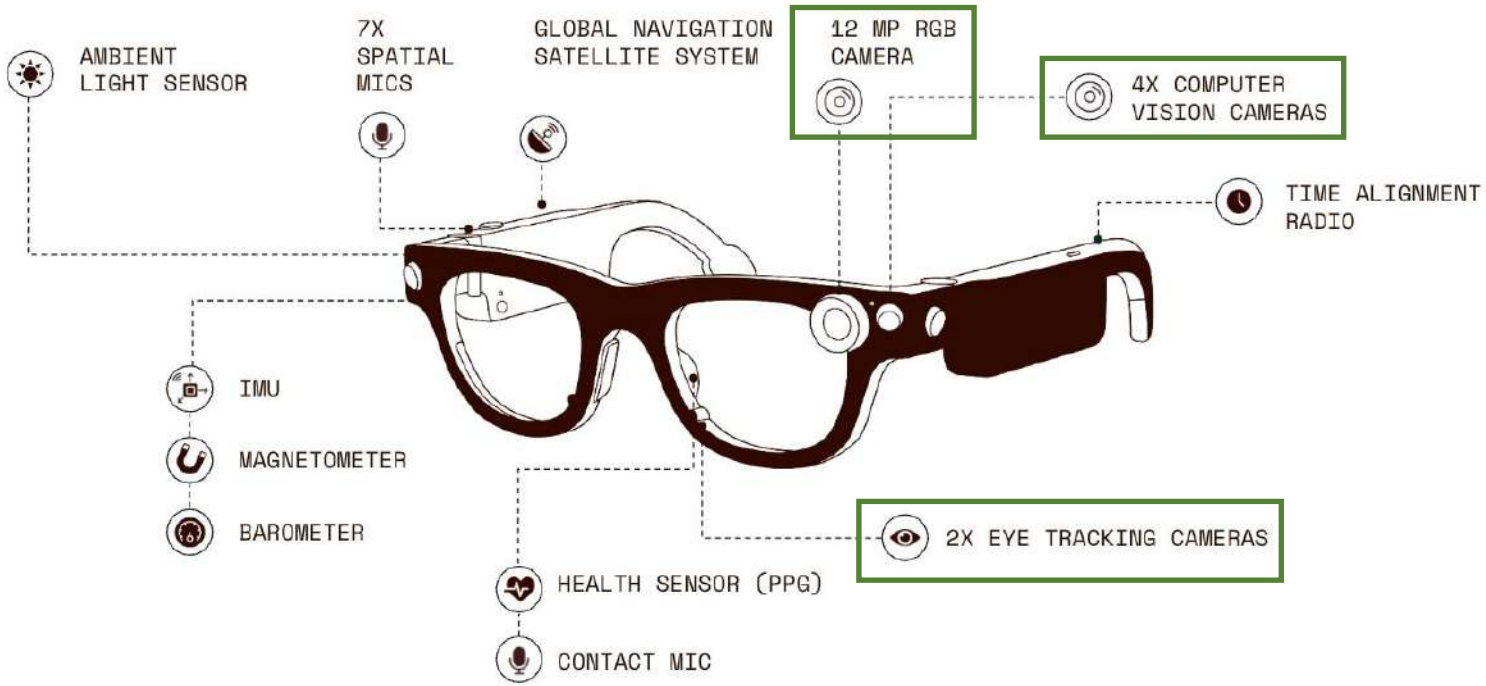
Smart Glasses

Energy Efficiency is key!

→ generative reasoning → multimodal perception → persistent context

Rich, Heterogeneous Sensing

Multiple sensors enable visual and contextual awareness



VISION

12 MP RGB camera

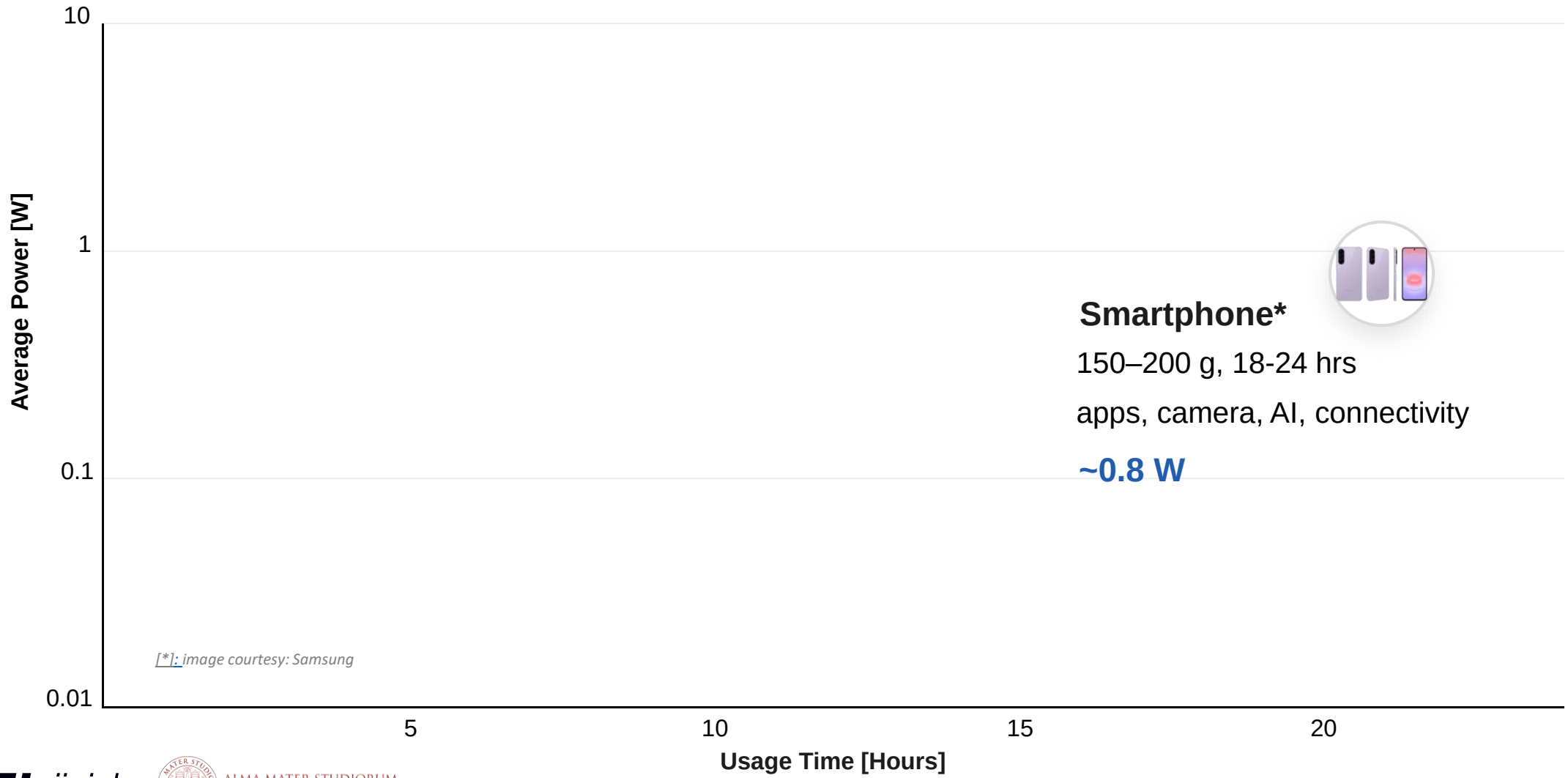
4× CV cameras

2× eye-tracking

Image adapted from Meta Project Aria Gen 2 documentation; background color modified.

Efficiency Requirements for Wearable Intelligence

From smartphones to XR headsets to smart glasses



Efficiency Requirements for Wearable Intelligence

From smartphones to XR headsets to smart glasses

XR headset*

500–800 g, 2-3 hrs

graphics, vision, tracking

~7 W

Smartphone

~0.8 W, 18 – 24 hrs

[*]: image courtesy: Apple Inc.

Average Power [W]

10

1

0.1

0.01

5

10

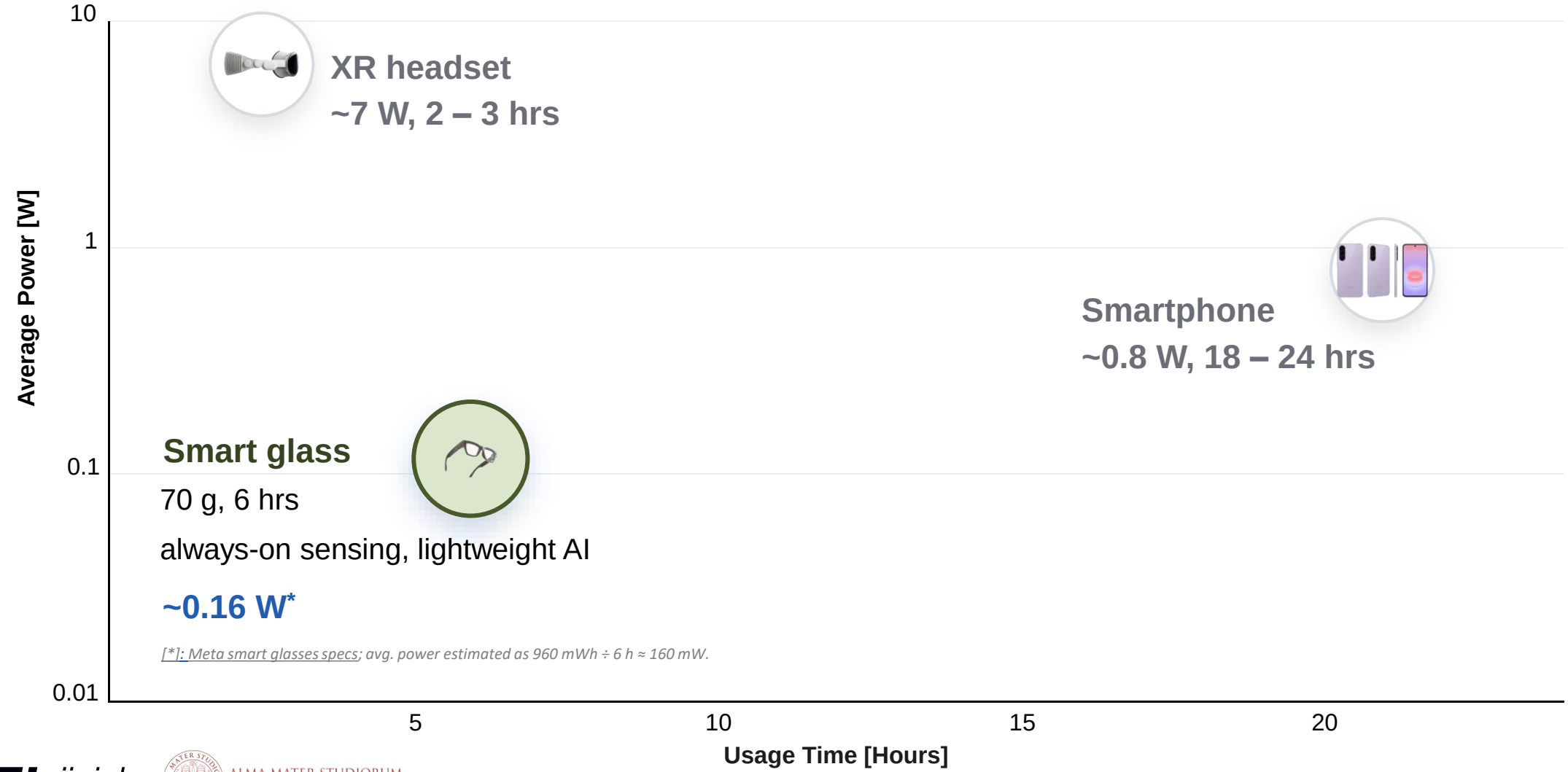
15

20

Usage Time [Hours]

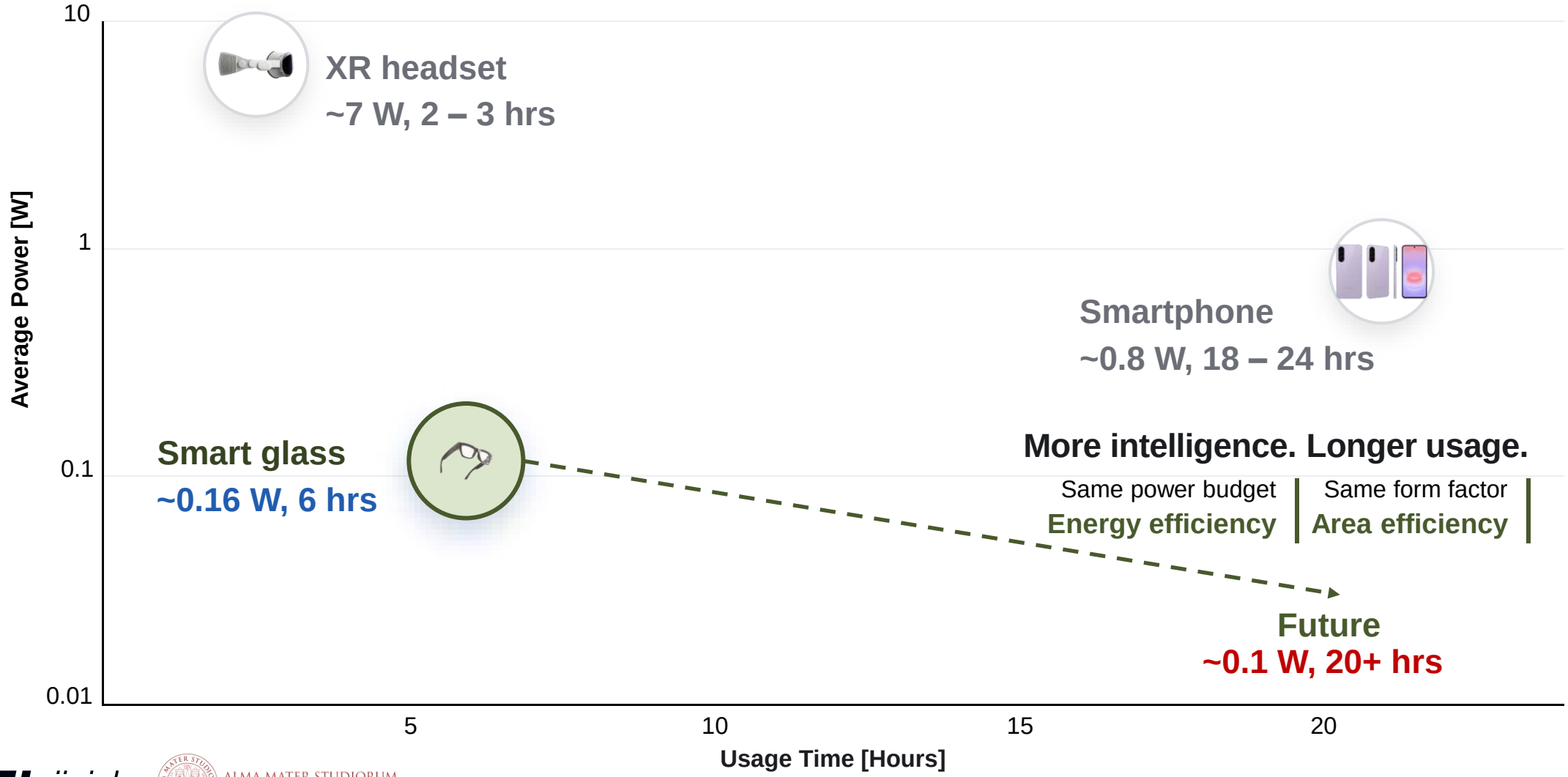
Efficiency Requirements for Wearable Intelligence

From smartphones to XR headsets to smart glasses



Efficiency Requirements for Wearable Intelligence

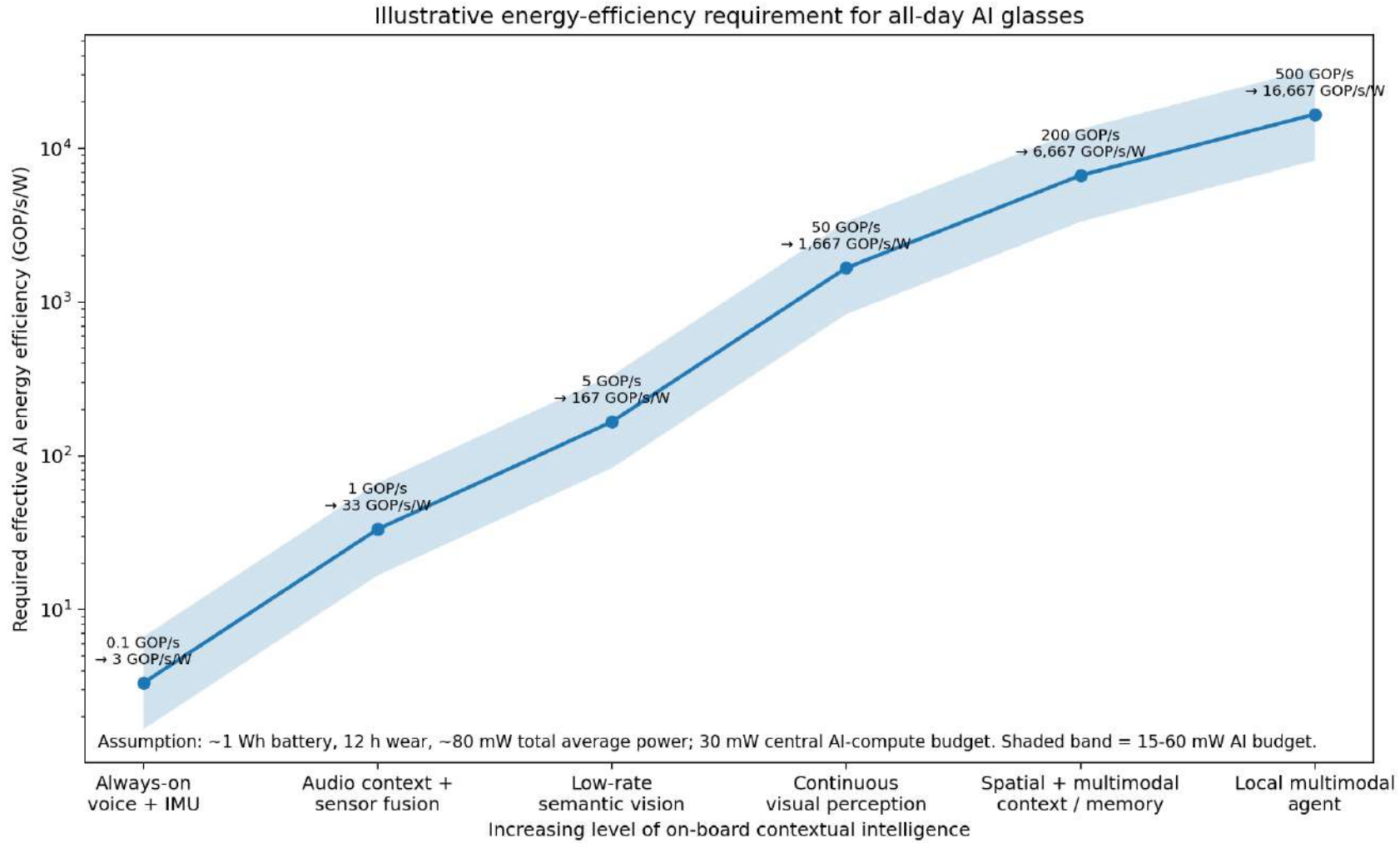
From smartphones to XR headsets to smart glasses



Efficiency Requirements for Wearable Intelligence

From perception to proactive agentic assistance

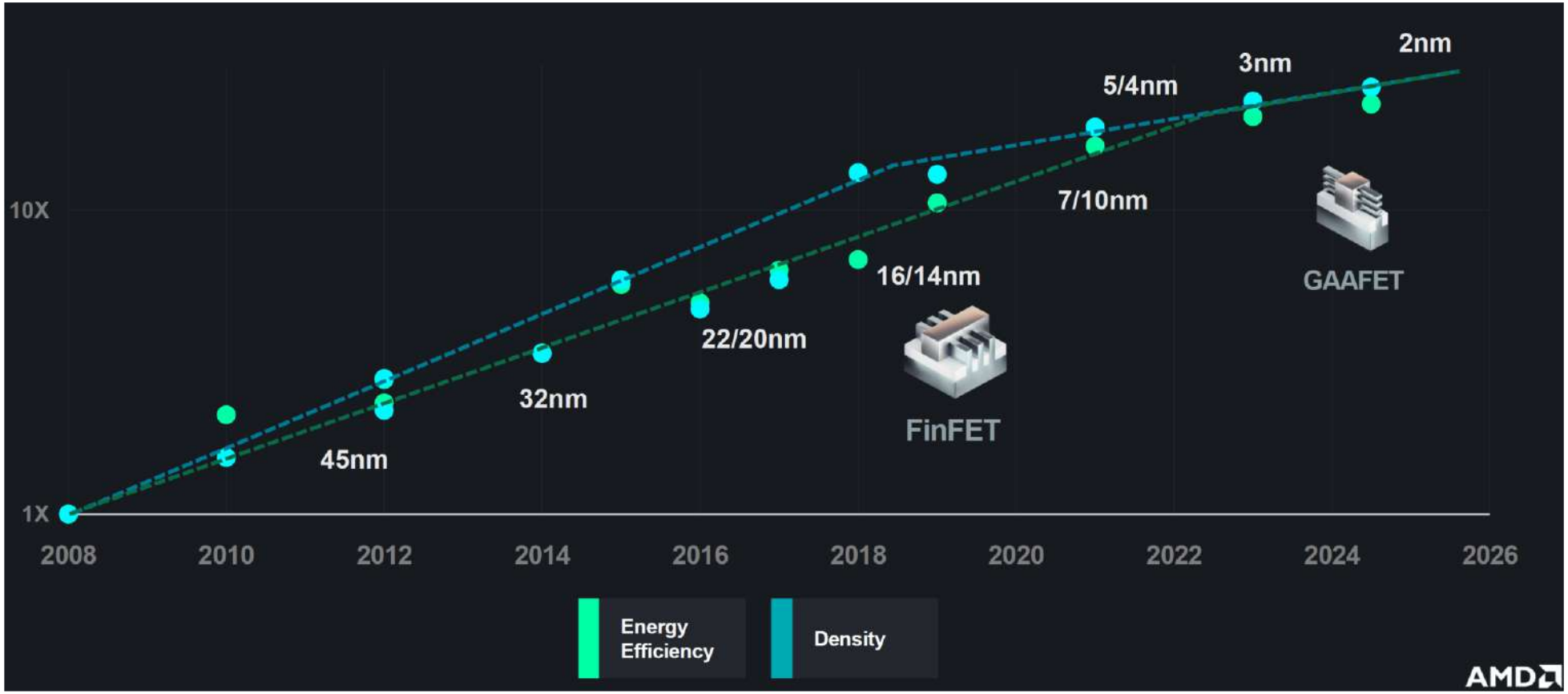
Energy efficiency →



Wearable Intelligence →

Generated via ChetGPT (SOL). Info from arXiv:2512.16045
T. Lee et al., Full System Architecture Modeling for Wearable Egocentric Contextual AI

Efficiency Challenge



[AMD HotChips24]

Efficient

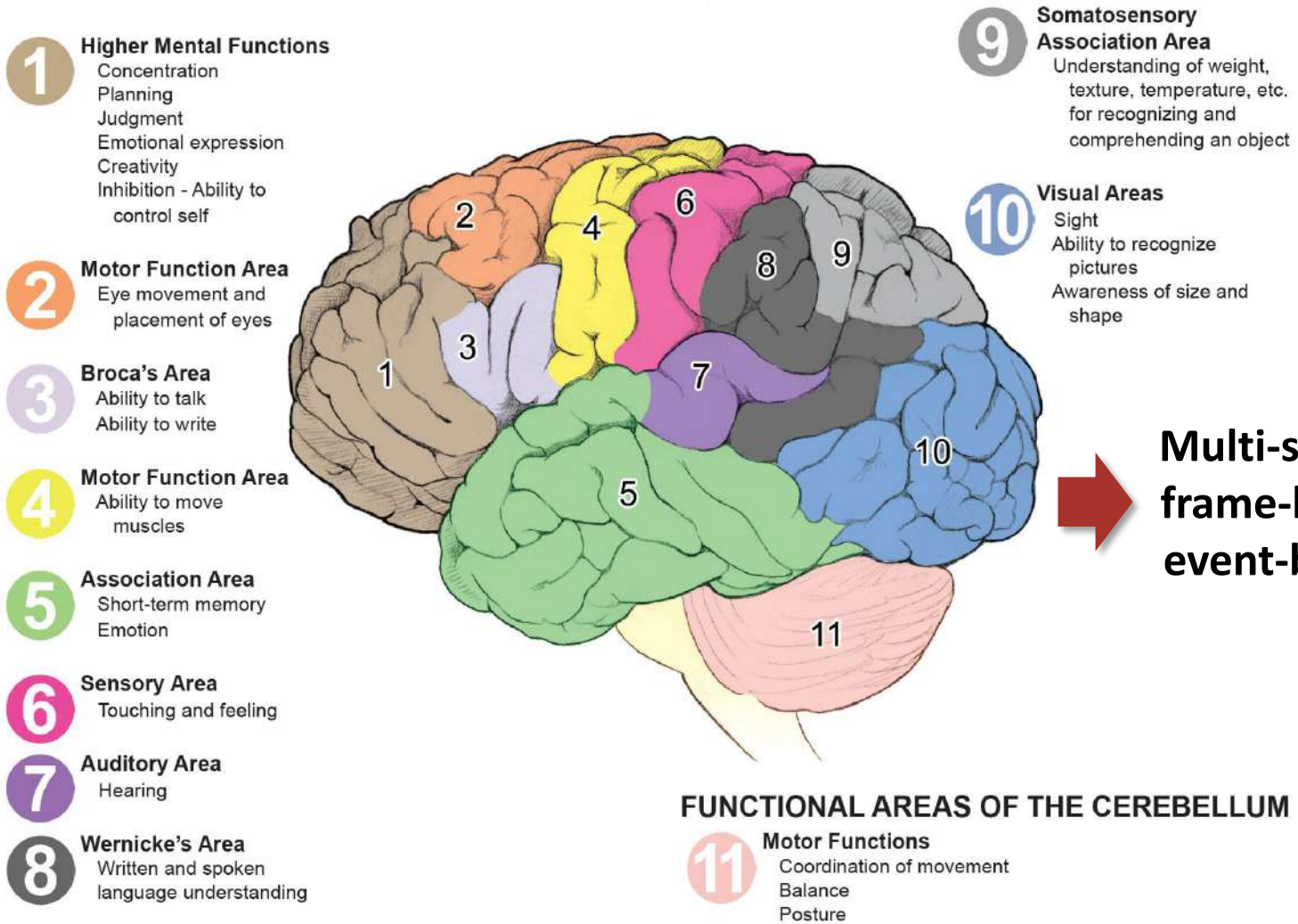
Model complexity
10× every ~2.5 years

Moore's Law
10x every 12 years!

Algorithm, Architecture, Design are key!

Efficiency via Heterogeneity: Multi-Domain Specialization

Brain-inspired: Multiple areas, different structure different function!

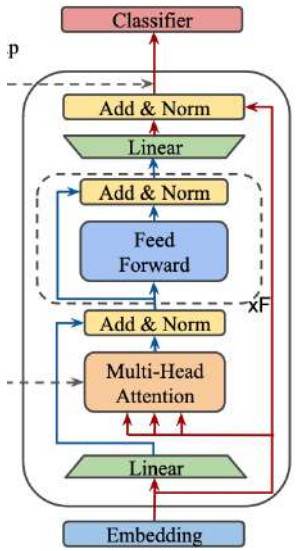


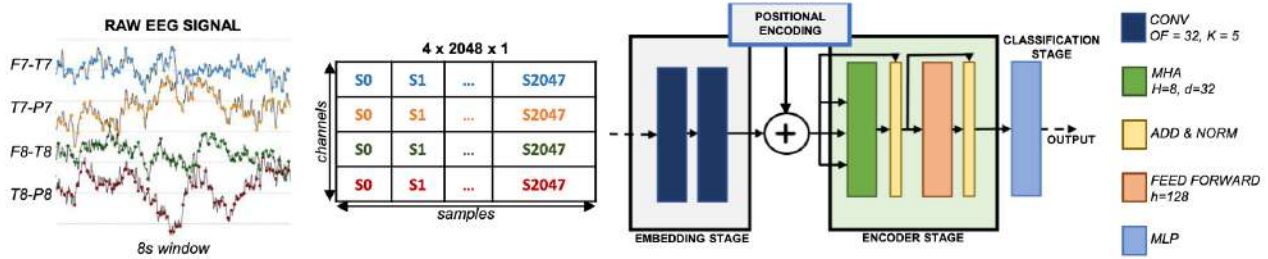
Multi-sensor frame-based event-based → **Perception** → **Multi-modal Agentic reasoning**

A Fast-Evolving Model Zoo

[Z. Sun et al.] MobileBERT

Encoder Transformer



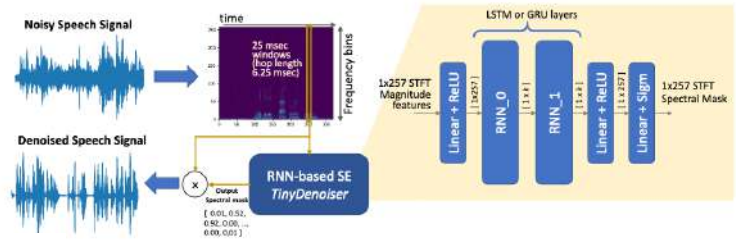


[P. Busia et al.] EEGFormer

Encoder Transformer

[M. Oquab et al.]

Encoder Transformer



[M. Rusci et al.] TinyDenoiser

Recurrent NN

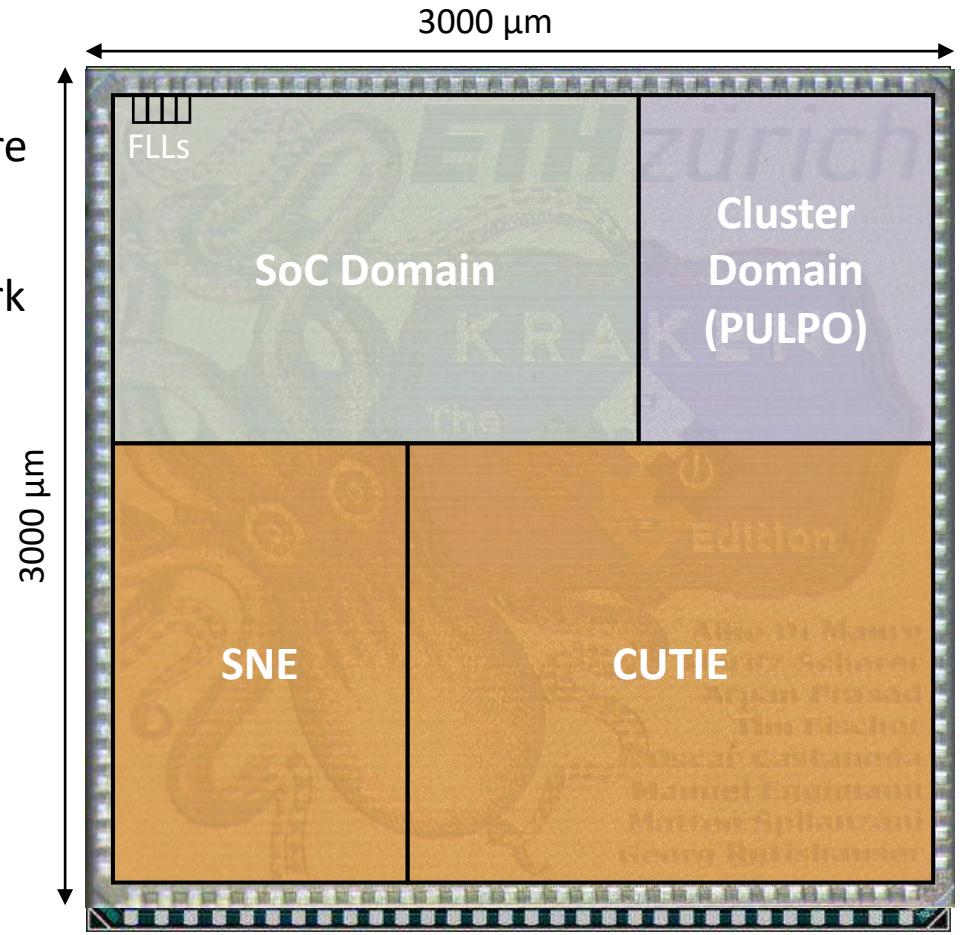
Auto-regressive Transformers

Walk on the specializazion ridge: specialize, but not too much!

Kraken: 22FDX SoC, Multi-Domain Specific Architecture

The ***Kraken***: an "Extreme Edge" Brain

- **RISC-V Cluster**
  8 Compute cores +1 DMA core
- **CUTIE**
  Dense ternary-neural-network accelerator
- **SNE**
  Energy-proportional spiking-neural-network accelerator



Technology	22 nm FDSOI
Chip Area	9 mm^2
SRAM SoC	1 MiB
SRAM Cluster	128 KiB
VDD range	0.55 V - **0.8 V**
Cluster Freq	**~370 MHz**
SNE Freq	**~250 MHz**
CUTIE Freq	**~140 MHz**

CUTIE: Perception from Frame Sensors

Ternary Activations
(2bits)

Ternary Weights
(2bits)

[Scherer et al. TCAD22]

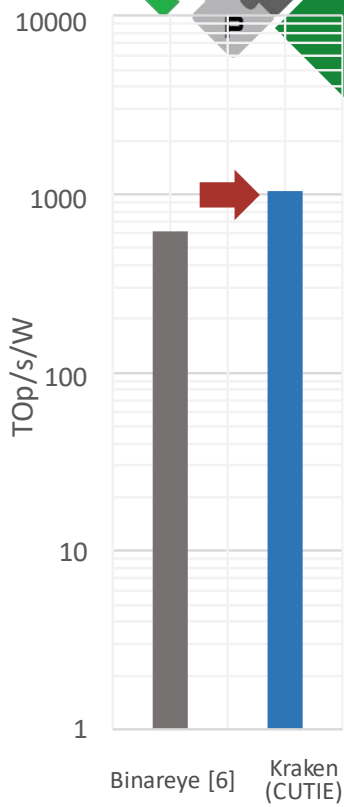
2 Popcount units

Pooling

Threshold

864 Ternary Multipliers

Output channel compute unit (OCU)



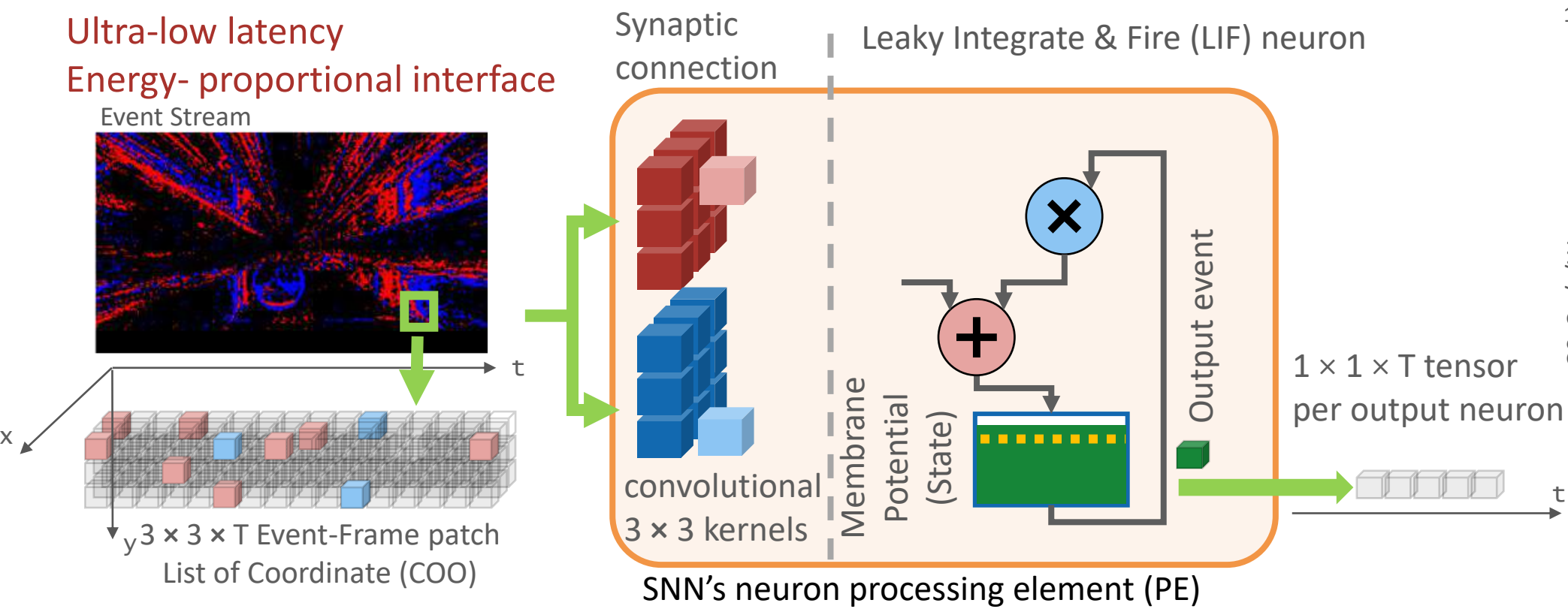
- **Completely Unrolled Ternary Neural Inference Engine**: K × K window, all input channels, cycle-by-cycle sliding
- One *Output Compute Unit* (OCU) computes one output activation per cycle!
- Zeros in weights and activations, spatial smoothness of activations reduce switching activity

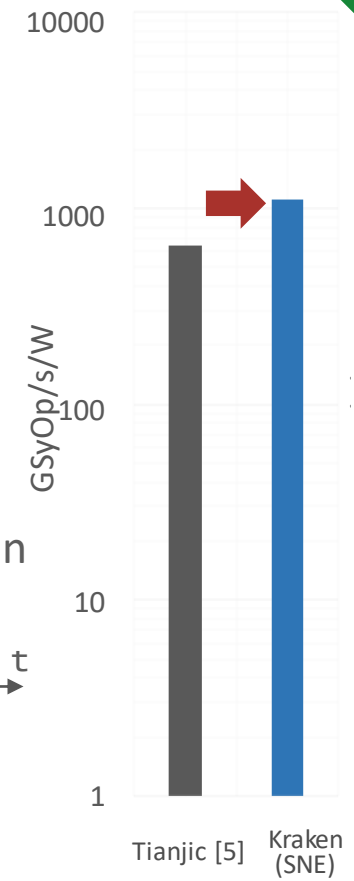
Aggressive quantization and full specialization

1fJ/MAC (1POP/s/W)
Ternary OPS

SNE: Perception from Event Sensors

Event Sensors – DVS camera
Ultra-low latency
Energy- proportional interface





A more complex dynamic than conventional DNNs neurons:

- Membrane Potential Accumulation/Activation 1× SynAcc = 1× 4b-ADD + 1× 8b-COMPARE
- Membrane Potential decay 1× SynDec = (1× 8b-MUL) + (1× 8b-MUL + 1× 8b-ADD)

1TSyOp/s/W

More Flexibility: Domain-Specific RV32 Core (PE)

RISC-V® Instruction set: open and extensible by construction (great!)

8-bit Convolution

Vanilla

N

```
addi   a0,a0,1
addi   t1,t1,1
addi   t3,t3,1
addi   t4,t4,1
lbu    a7,-1(a0)
lbu    a6,-1(t4)
lbu    a5,-1(t3)
lbu    t5,-1(t1)
mul    s1,a7,a6
mul    a7,a7,a5
add    s0,s0,s1
mul    a6,a6,t5
add    t0,t0,a7
mul    a5,a5,t5
add    t2,t2,a6
add    t6,t6,a5
bne    s5,a0,1c000bc
```

RISC-V core

Specialized for AI → Mixed precision SIMD (16-2bit)

N/4

```
Init NN-RF (outside of the loop)
lp.setup
pv.nnsdotup.h  s0,ax1,9
pv.nnsdotsp.b  s1, aw2, 0
pv.nnsdotsp.b  s2, aw4, 2
pv.nnsdotsp.b  s3, aw3, 4
pv.nnsdotsp.b  s4, ax1, 14
end
```

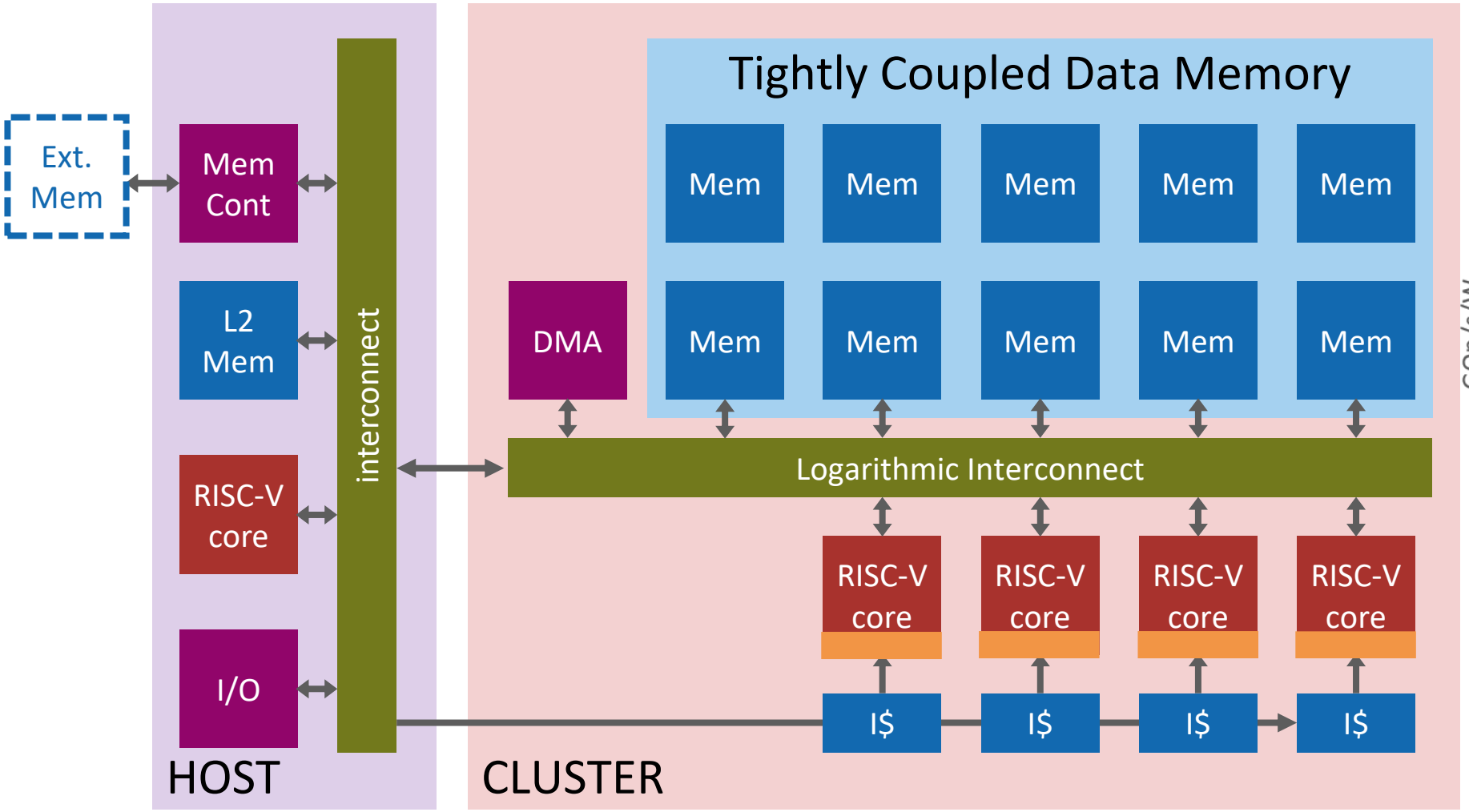
RISC-V core

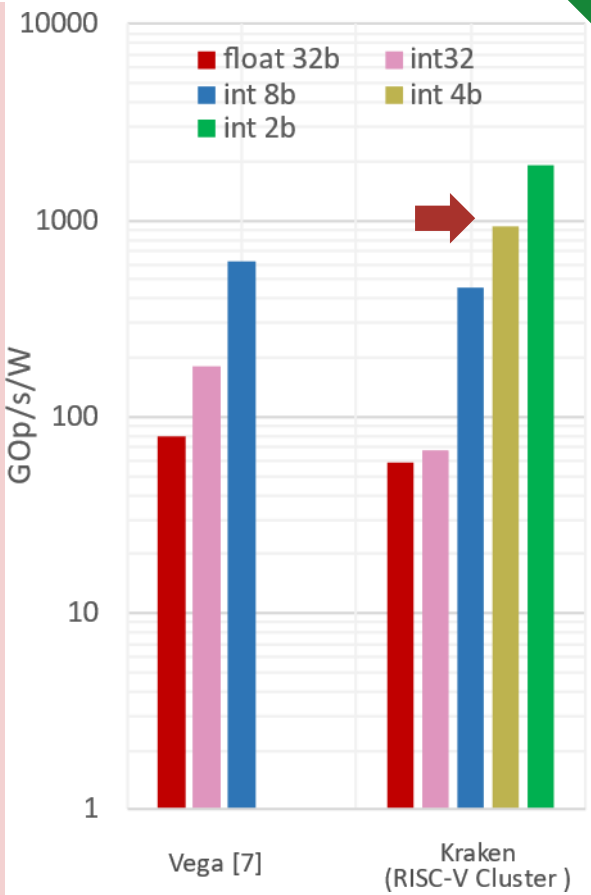
15x less instructions than Vanilla

90%+ ALU Utilization

Specialization Cost: Power, Area: 1.5×↑ Time 15×↓ → E = PT 10× ↓

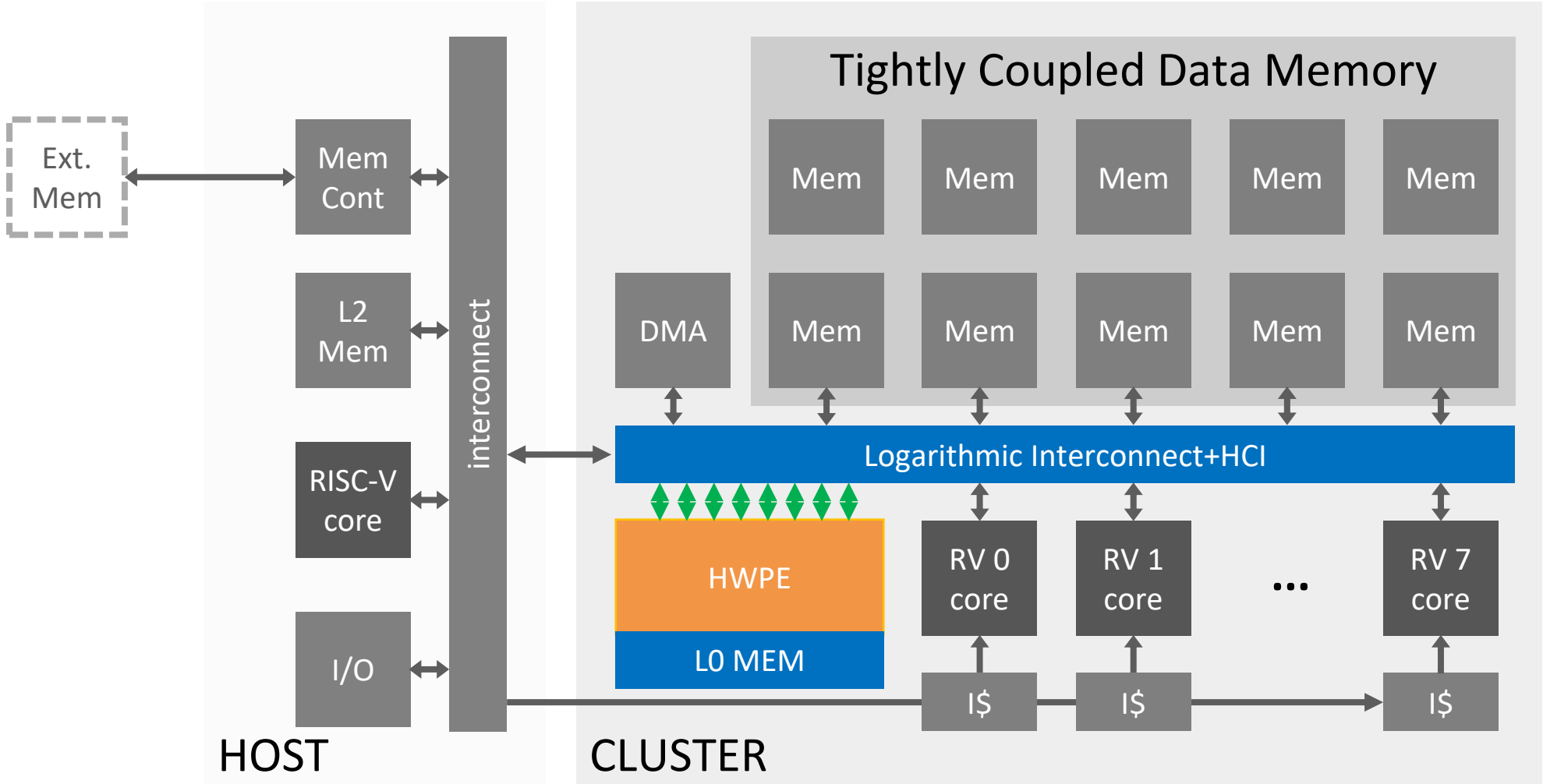
A Cluster of Domain Specific Cores





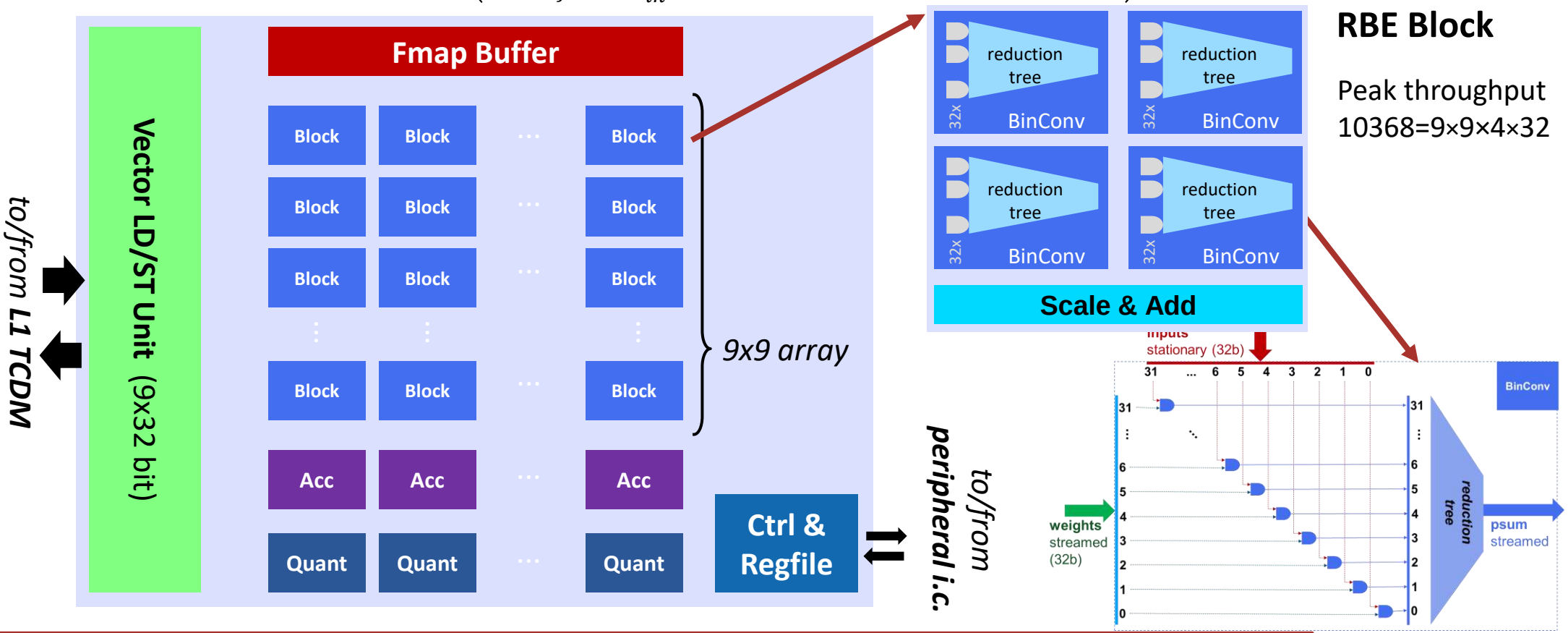
1TOP/s/W 2b/4b OPS

Boosting Efficiency: Tightly-coupled Accelerators



HWPE: Reconfigurable Binary Engine

$$\mathbf{y}(k_{out}) = \boldsymbol{quant}\left(\sum_{i=0..M}\sum_{j=0..N}\sum_{k_{in}} 2^i 2^j \left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x_{bin}}(k_{in})\right)\right)$$



Energy efficiency 10-20× (0.1pJ/OP) w.r.t. SW on cluster @same accuracy

Specialization in Perspective: Amdahl's Law

Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-Efficient RV Core → **20pJ (8bit)**

↓

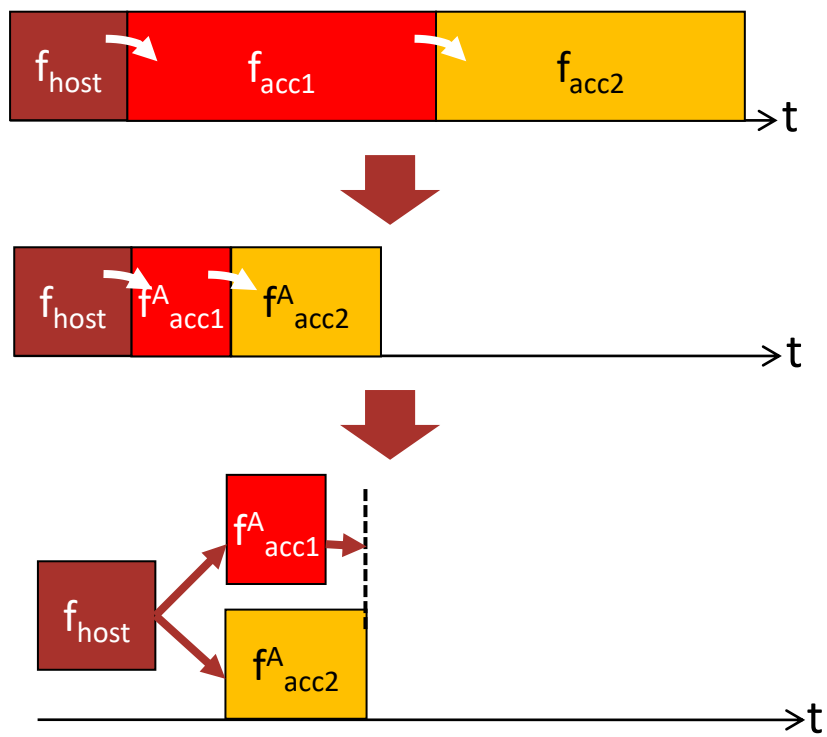
ISA-based 10-20x →**1pJ (4bit)**

↓

Tightly coupled 10-20x → **100fJ (4bit)**

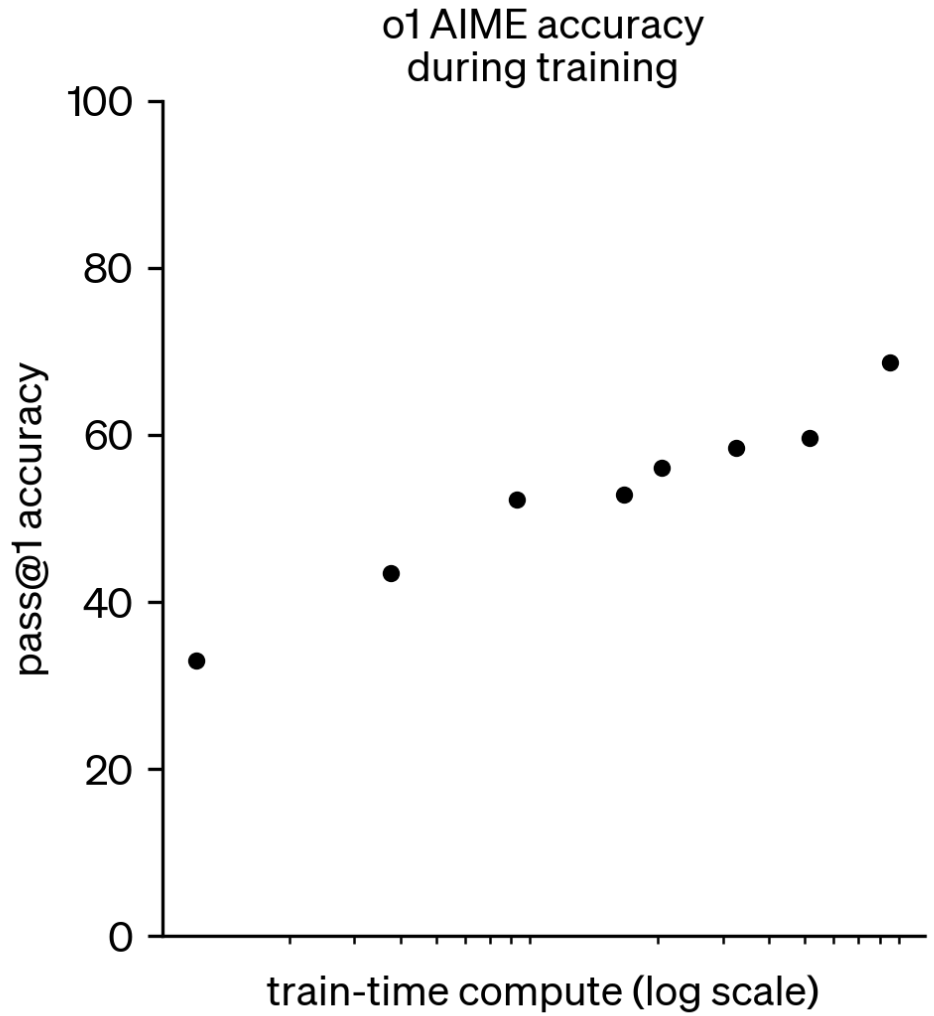
↓

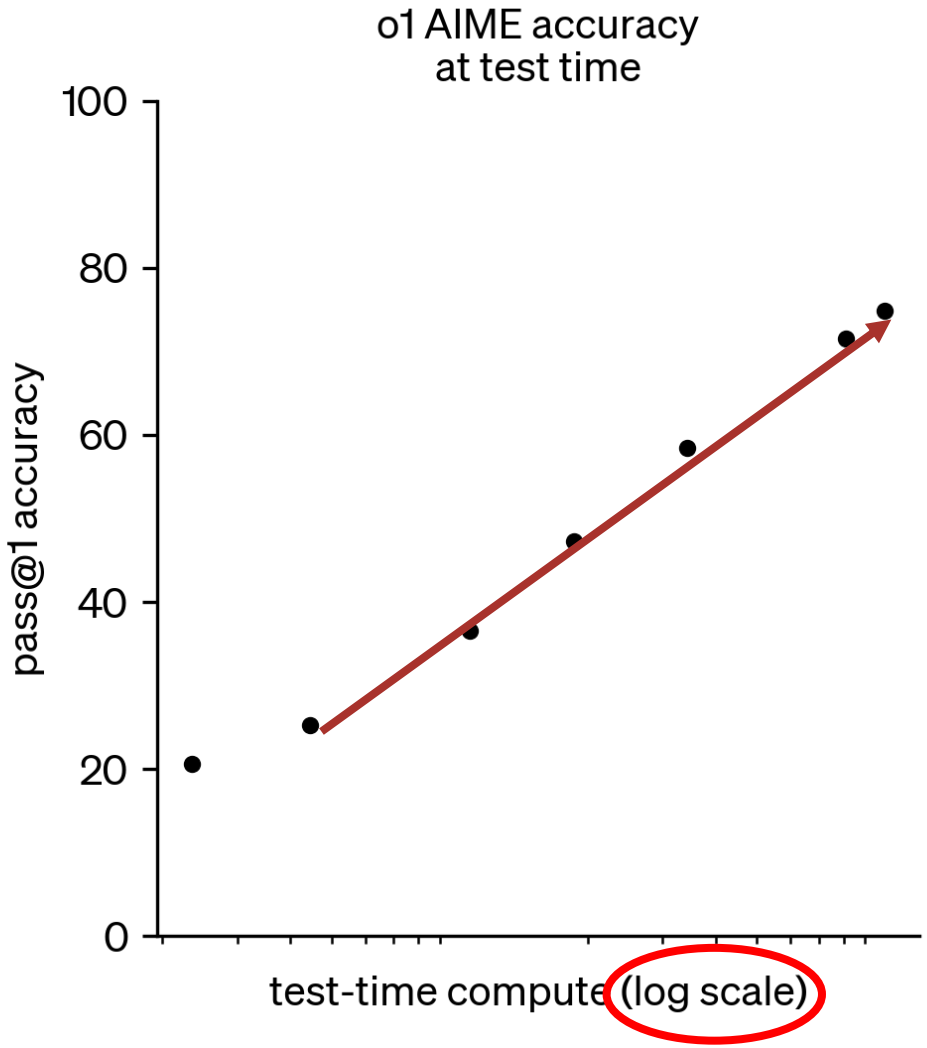
Decoupled 100x →**1fJ (ternary)**



Specialization game: efficiency vs. utilization Tradeoff

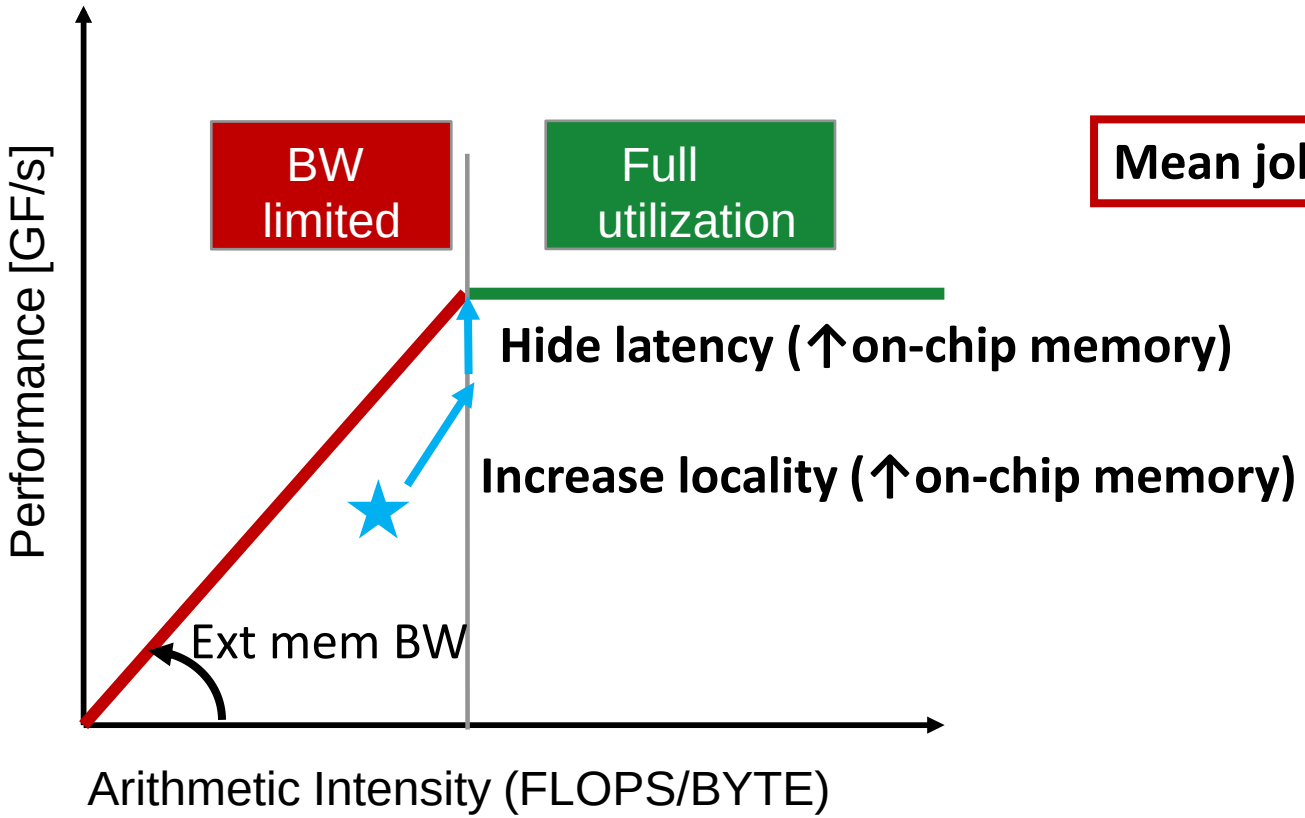
Generative AI Scaling Laws





[openai.com/index/learning-to-reason-with-llms/]

Scalability in Perspective: Patterson (roofline), Little Laws



Mean jobs in system = arrival rate x mean response time

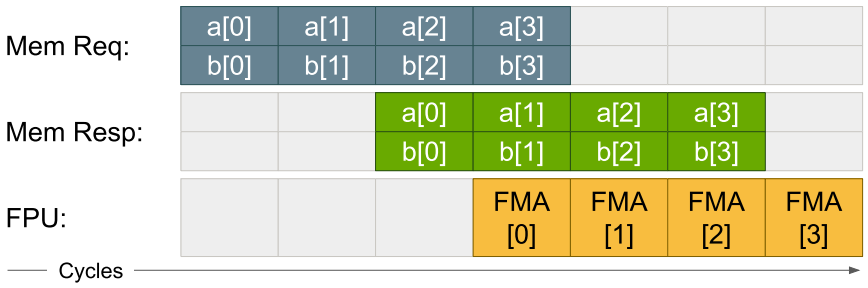
BufferSize (B) = Bandwidth (B/s) x Latency (s)

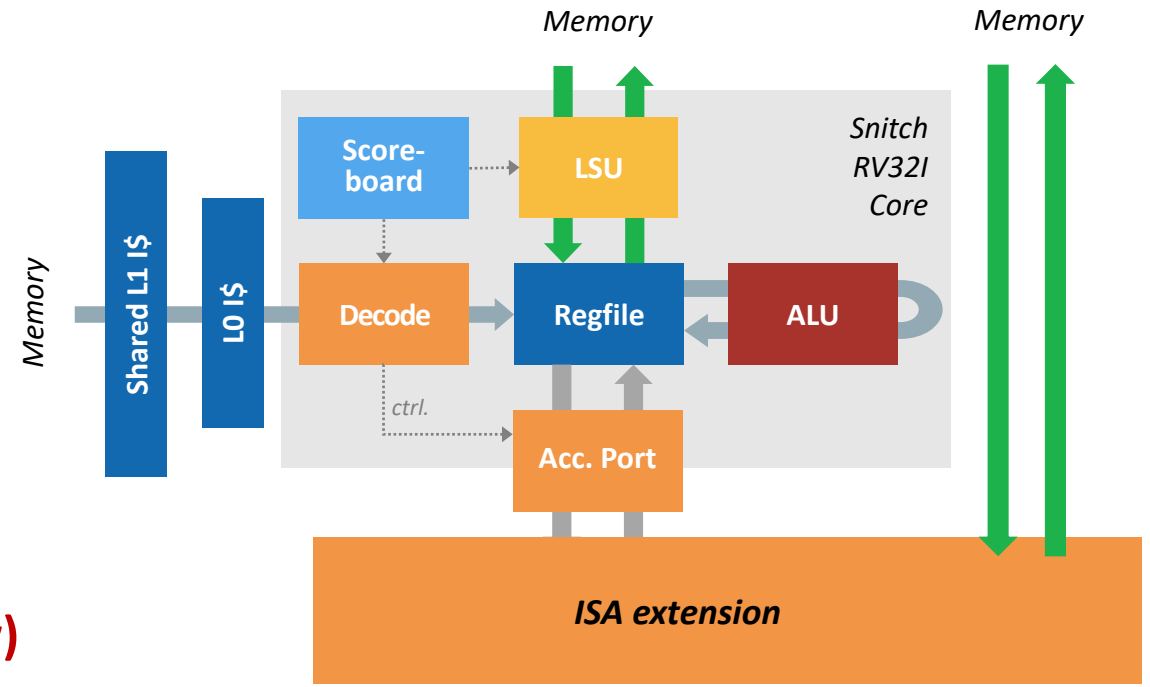
Eg. 1KB/ns x 100 (ns) → 100KB

Need on-chip Buffering, and latency tolerant Architecture

Latency Tolerance: The Snitch Core

- ***Snitch*: tiny (20KGE), extensible RV core**
  - *Extensible* through **accelerator port**

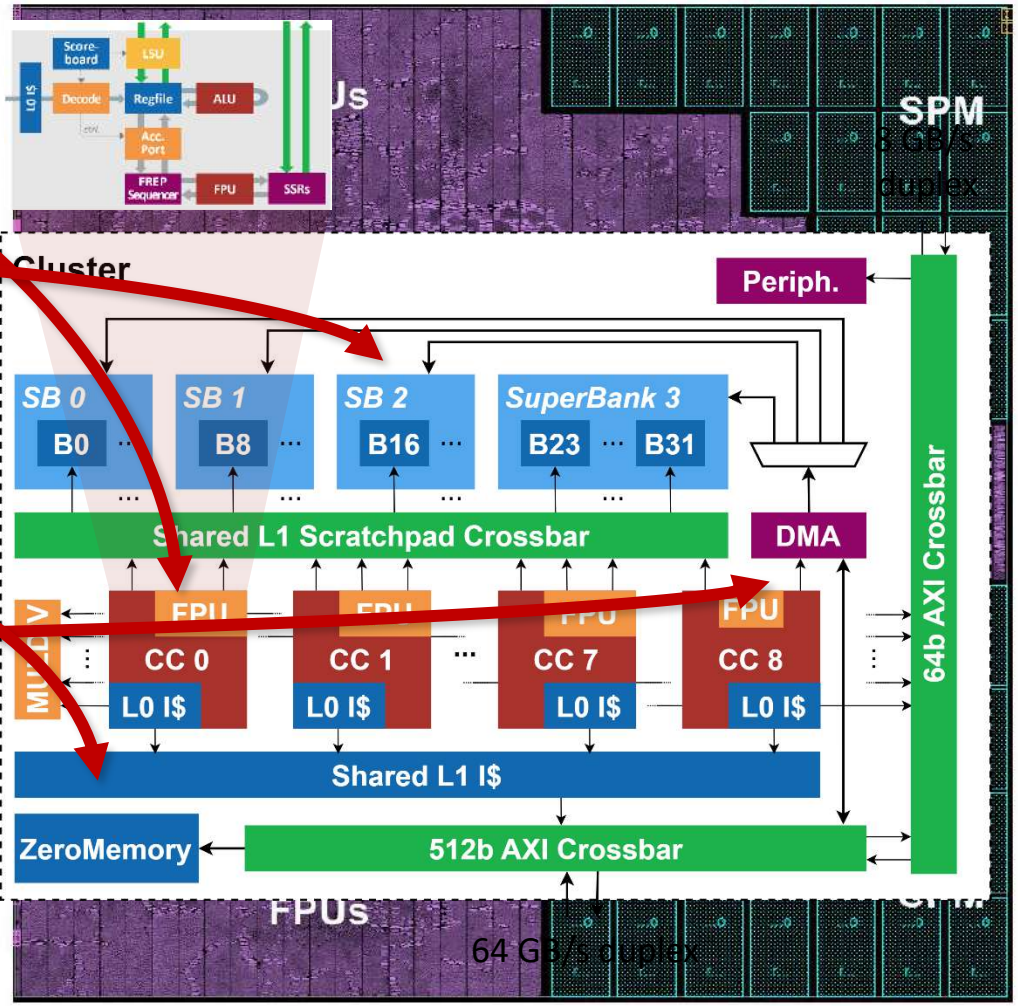




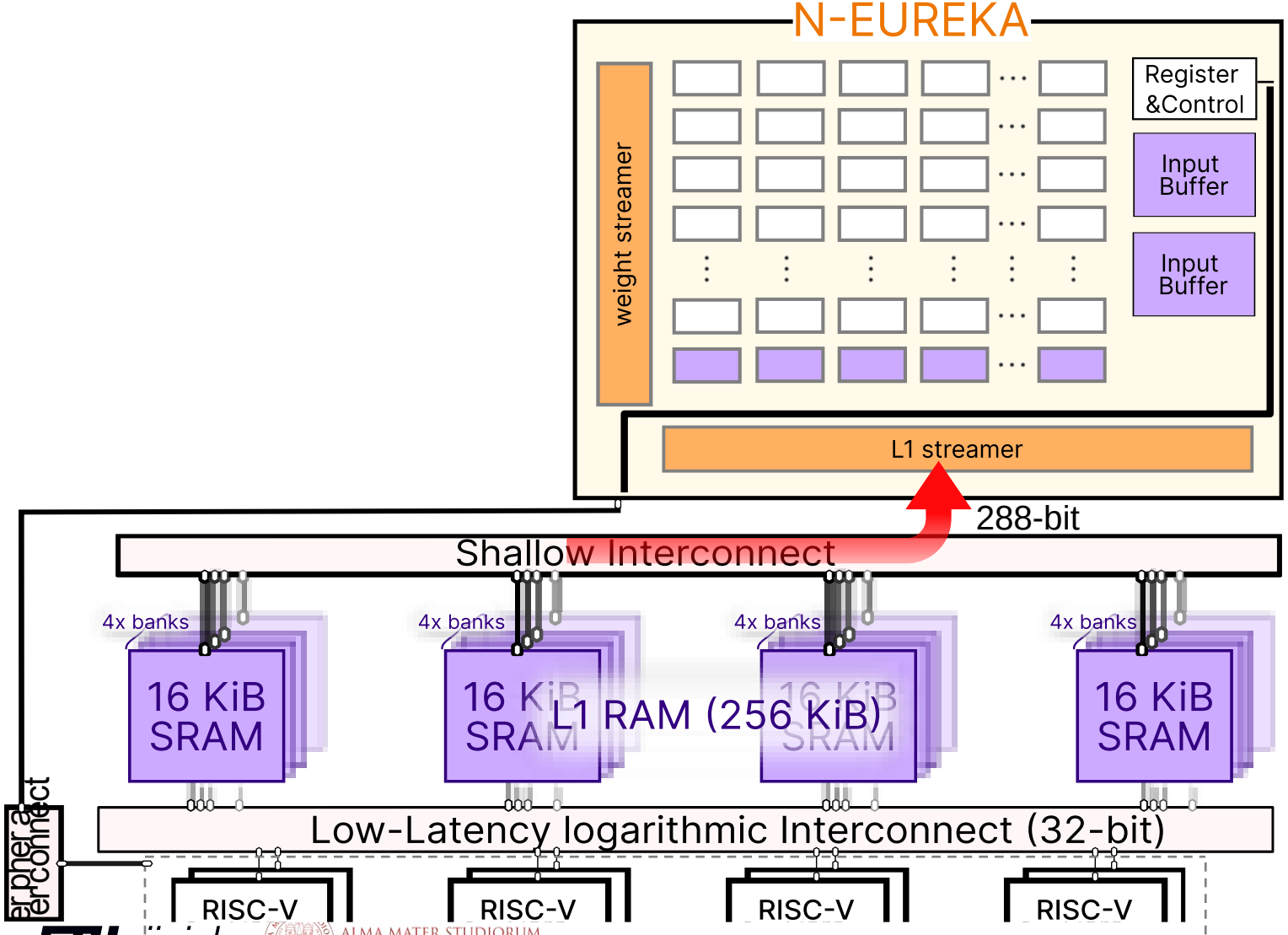
  - *Latency-tolerant* through scoreboard+ld/st queue
    → can issue ~10 non-blocking memOPs
  - **Tolerates 10 cycles of memory latency (Little's law)**

Latency Tolerance: The Smart Data Mover

- **Snitch compute cores**
  - SIMD, Vector, Matrix etension
  - Latency tolerant
- **L1 Shared TCDM**
  - Less than 10 cycles latency shared scratchpad
- **Shared I-cache and peripherals**
- **DMA engine**
  - 512b interface to interconnect
  - HW support for autonomous multi dim. transfers
- **Shared DMA (10% overhead) for latency tolerance of L2+→L1: 100s vs 10s cycles**



On-Chip Buffering: Bandwidth and Capacity



N-Eureka (Specialization)

CNN workloads (3 × 3, DW, PW)

2 – 8-bit weight, 8-bit input, 32-bit accumulation

parametric PEs (288 × 1 × 8-bit multipliers)

input buffer → data reuse

memory-mapped config (2× contexts)

L1 streamer (288-bit/cycle)

weight streamer → utilization ↑

Shallow interconnect (wide, contiguous)

L1 memory (multi-banked, interleaved)

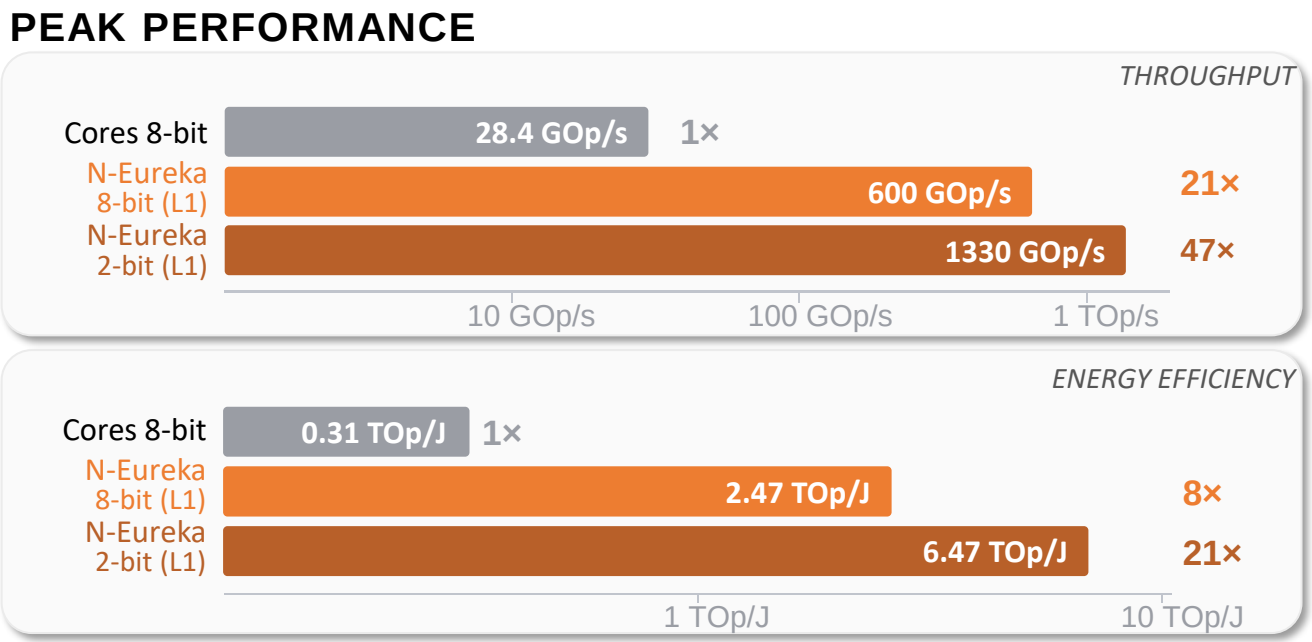
Logarithmic interconnect (narrow, flexible)

DSP-enhanced cores (octa-core)

On-Chip Buffering: Bandwidth and Capacity

Does **peak performance** translate to ***end-to-end*** **system performance**?

- Specialization → N-Eureka CNN engine
- Shallow interconnect (wide, contiguous)
- L1 memory (multi-banked, interleaved)
- Logarithmic interconnect (narrow, flexible)
- DSP-enhanced cores (octa-core)



On-Chip Buffering: Bandwidth and Capacity

Beyond compute:
off-chip memory access limits both latency and energy

Specialization → N-Eureka CNN engine

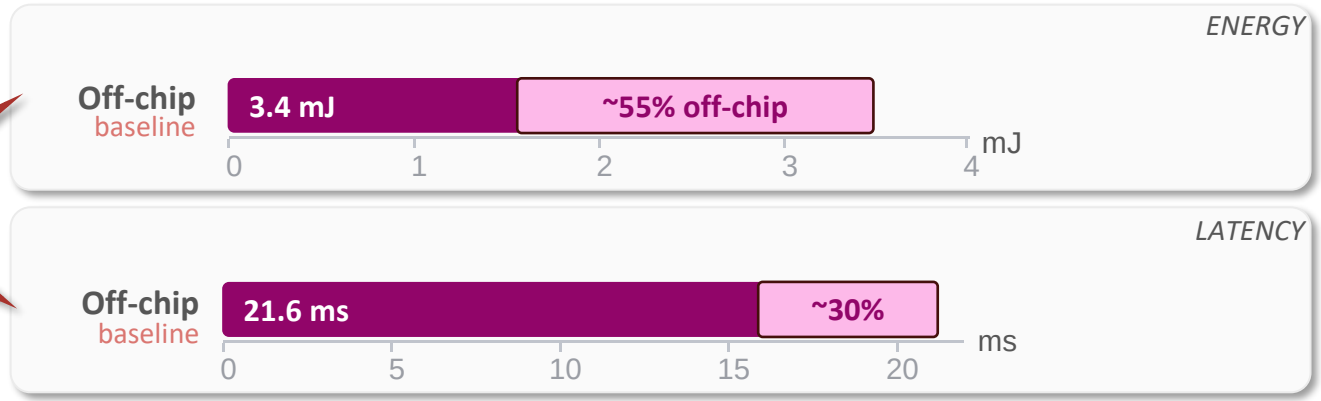
Shallow interconnect (wide, contiguous)

L1 memory (multi-banked, interleaved)

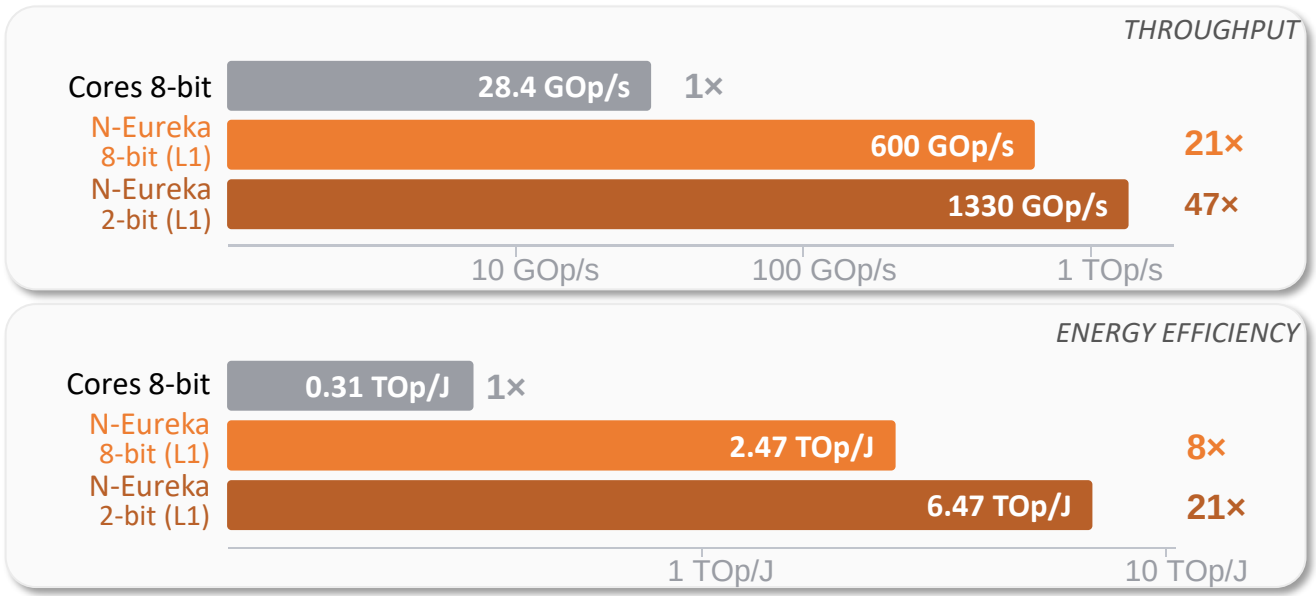
Logarithmic interconnect (narrow, flexible)

DSP-enhanced cores (octa-core)

END TO END PERFORMANCE



PEAK PERFORMANCE

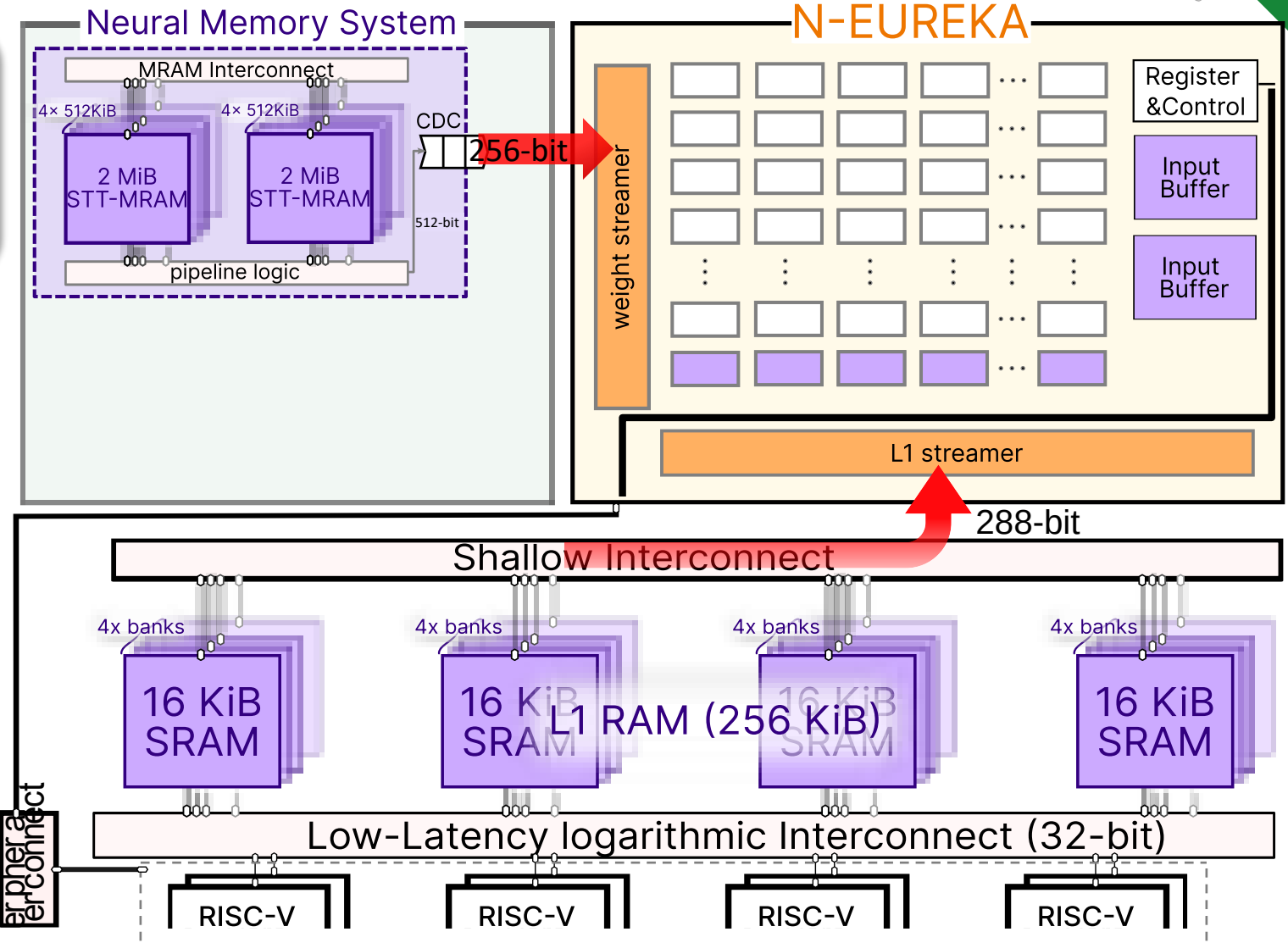


At-Memory Computing

Specialize! Not only Compute, but also **Interconnect, Memory**

MRAM (4 MiB):

- **non-volatile** → retain data
- **high density** → store more data
- extremely slow write (~10×)
- slow read (~2×)

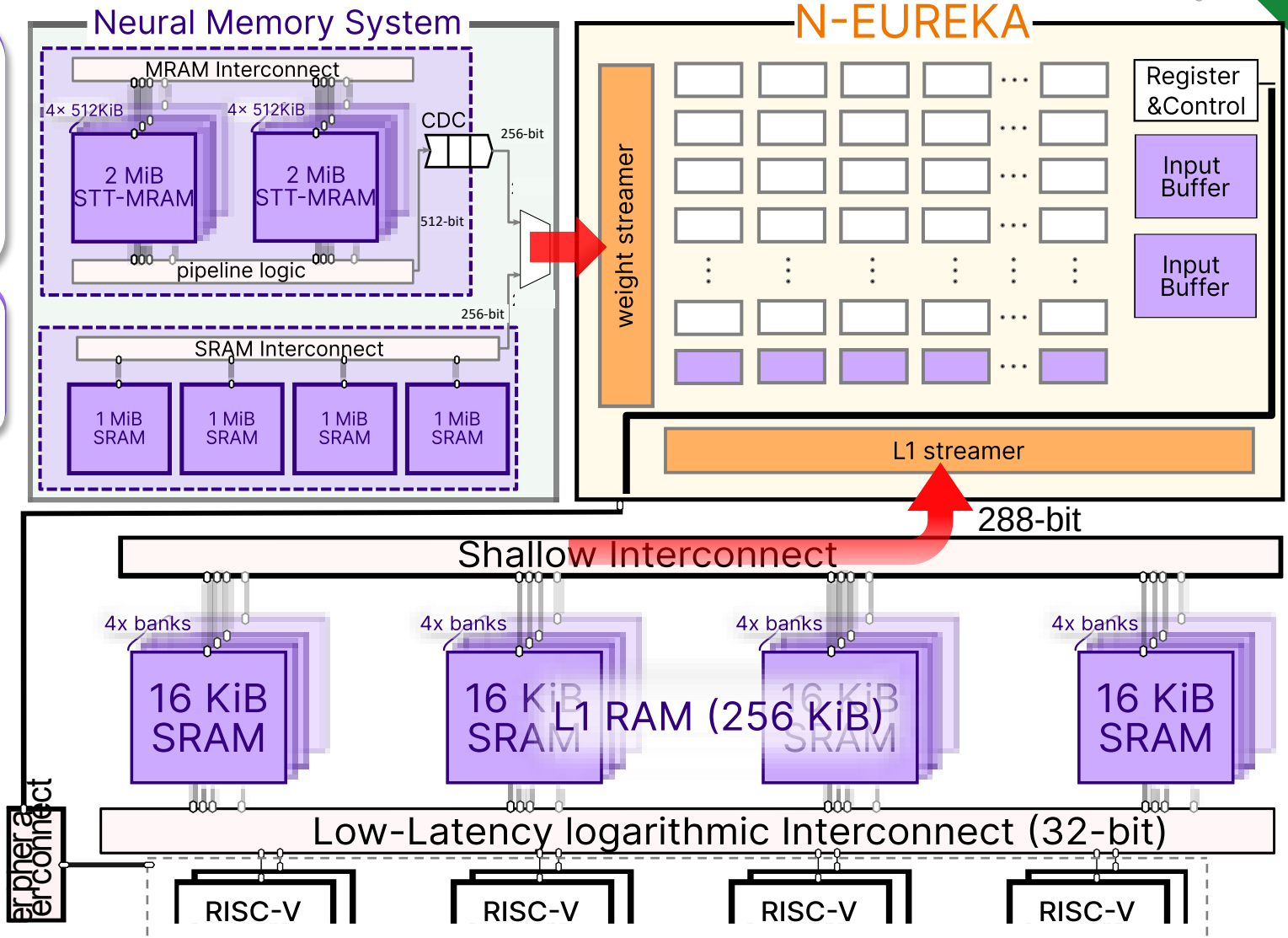


At-Memory Computing

Specialize! Not only Compute, but also **Interconnect, Memory**

MRAM (4 MiB):
- **non-volatile** → retain data
- **high density** → store more data
- extremely **slow write (~10×)**
- **slow read (~2×)**

SRAM (4 MiB):
- dynamic **activation** storage
- optionally **weight** storage



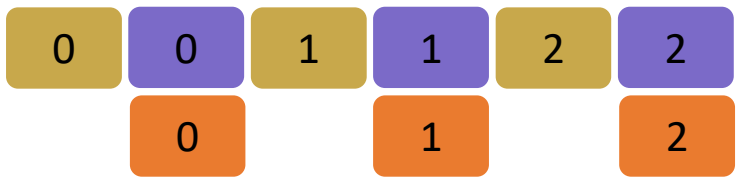
At-Memory Computing

Specialize! Not only Compute, but also **Interconnect, Memory**

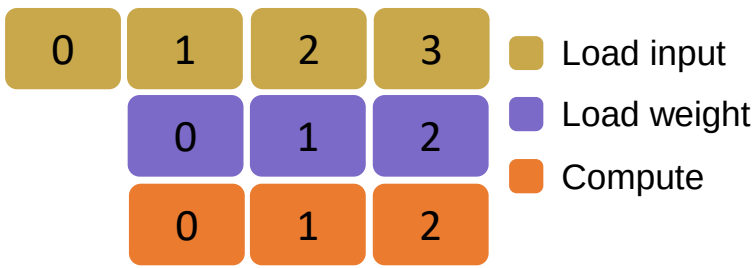
MRAM (4 MiB):
non-volatile, **high density** weight memory

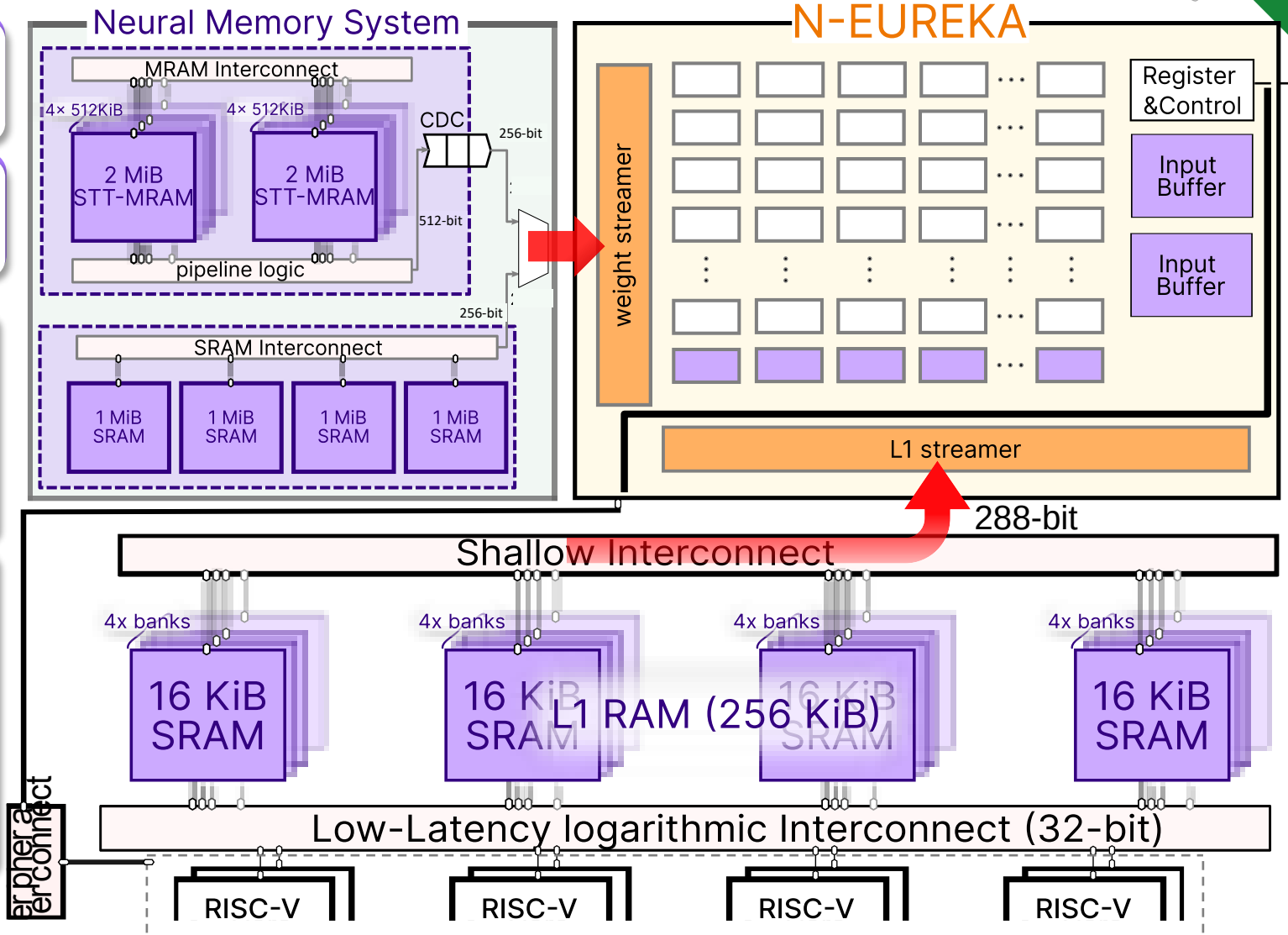
SRAM (4 MiB):
dynamic **activation,** optionally **weight**

Before



After faster



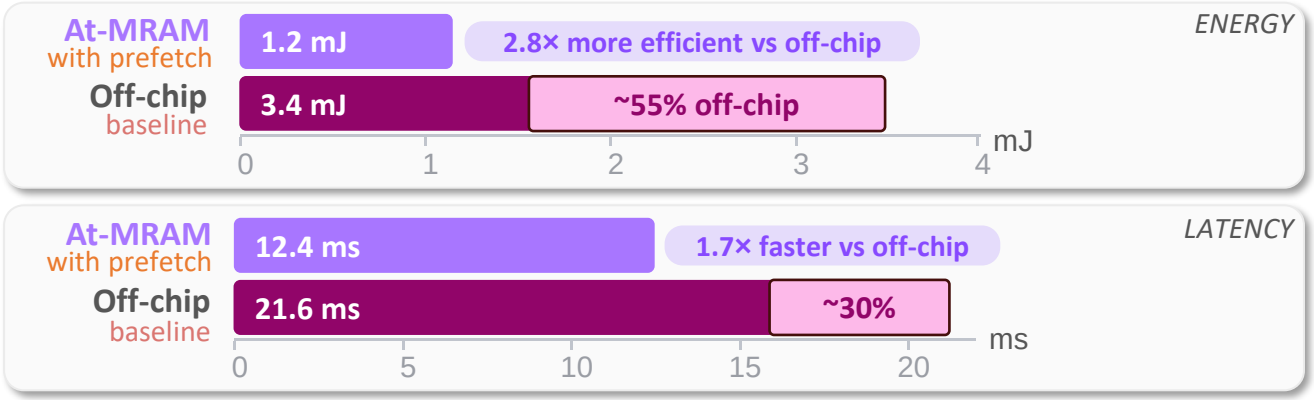


At-Memory Computing

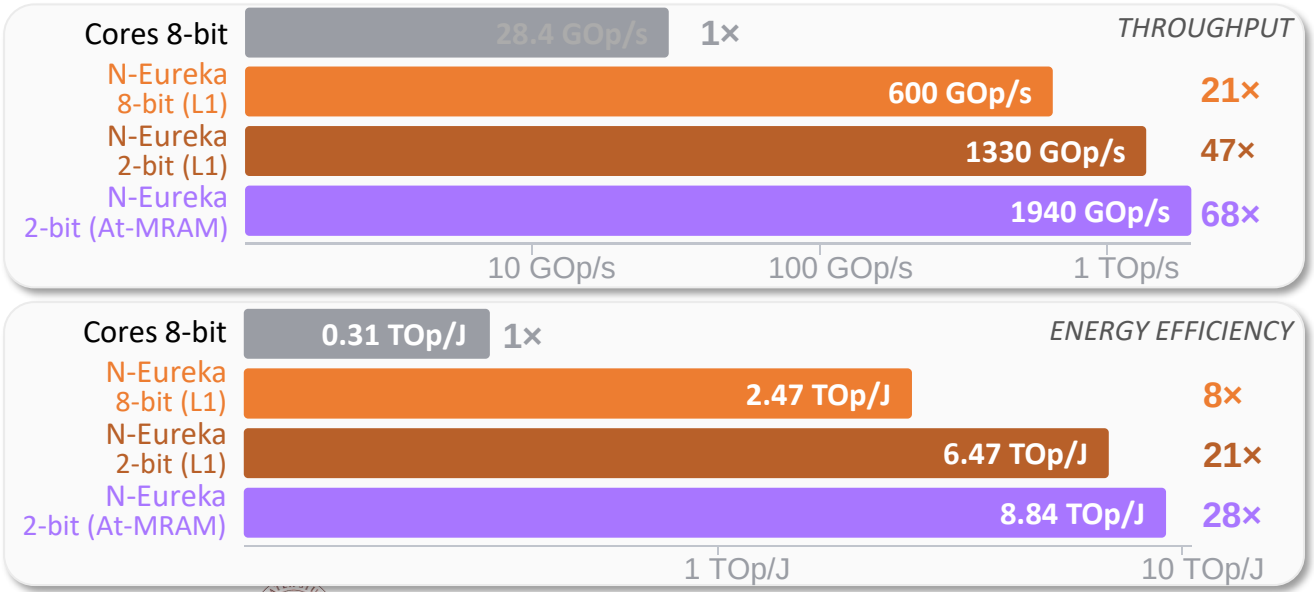
Specialize! Not only Compute, but also **Interconnect, Memory**

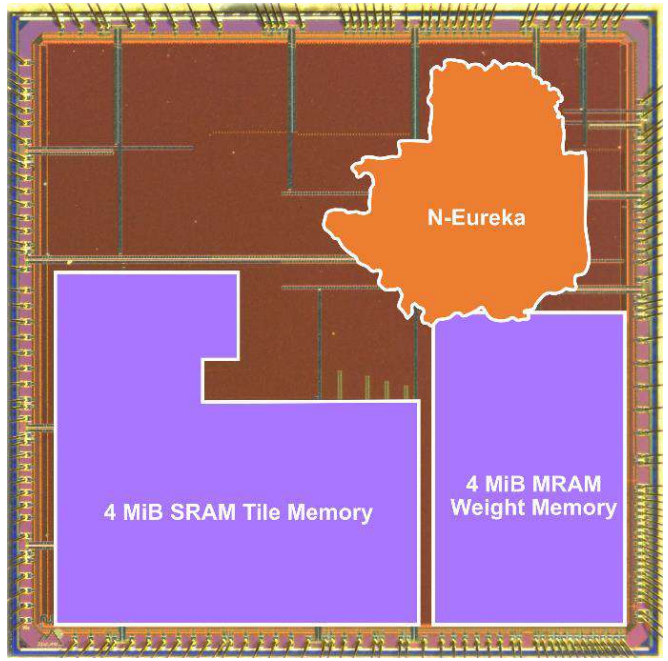
SIRACUSA

END-TO-END PERFORMANCE



PEAK PERFORMANCE





TECH	AREA	FREQUENCY	POWER
16 nm	16 mm^2	210 – 360 MHz	151 – 332 mW

Comparison with SoA (8-bit)

4.8×	1.3×	1.18×
throughput	energy efficiency	area efficiency

Flexible Domain-Specific Architecture

Multiple dimensions for Specialization

Extensions to processor cores

ISA extension

Same L0 (RF)

Tensor Engines

ISA extension or fast offload

Decoupled L0

Decoupled Accelerators

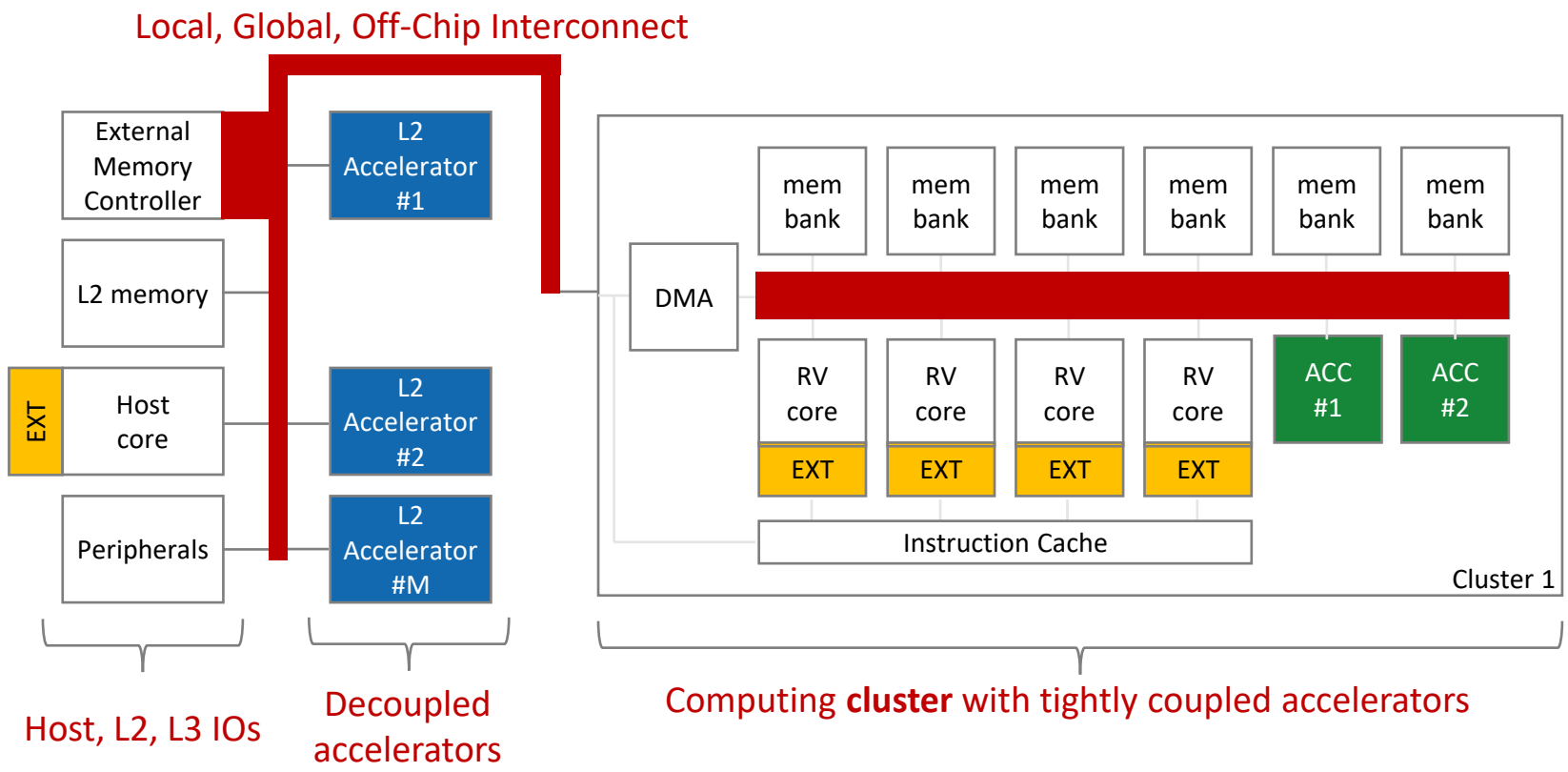
Explicit Coarse Grained offoad

Decoupled L1

Specialized Mem and Interco

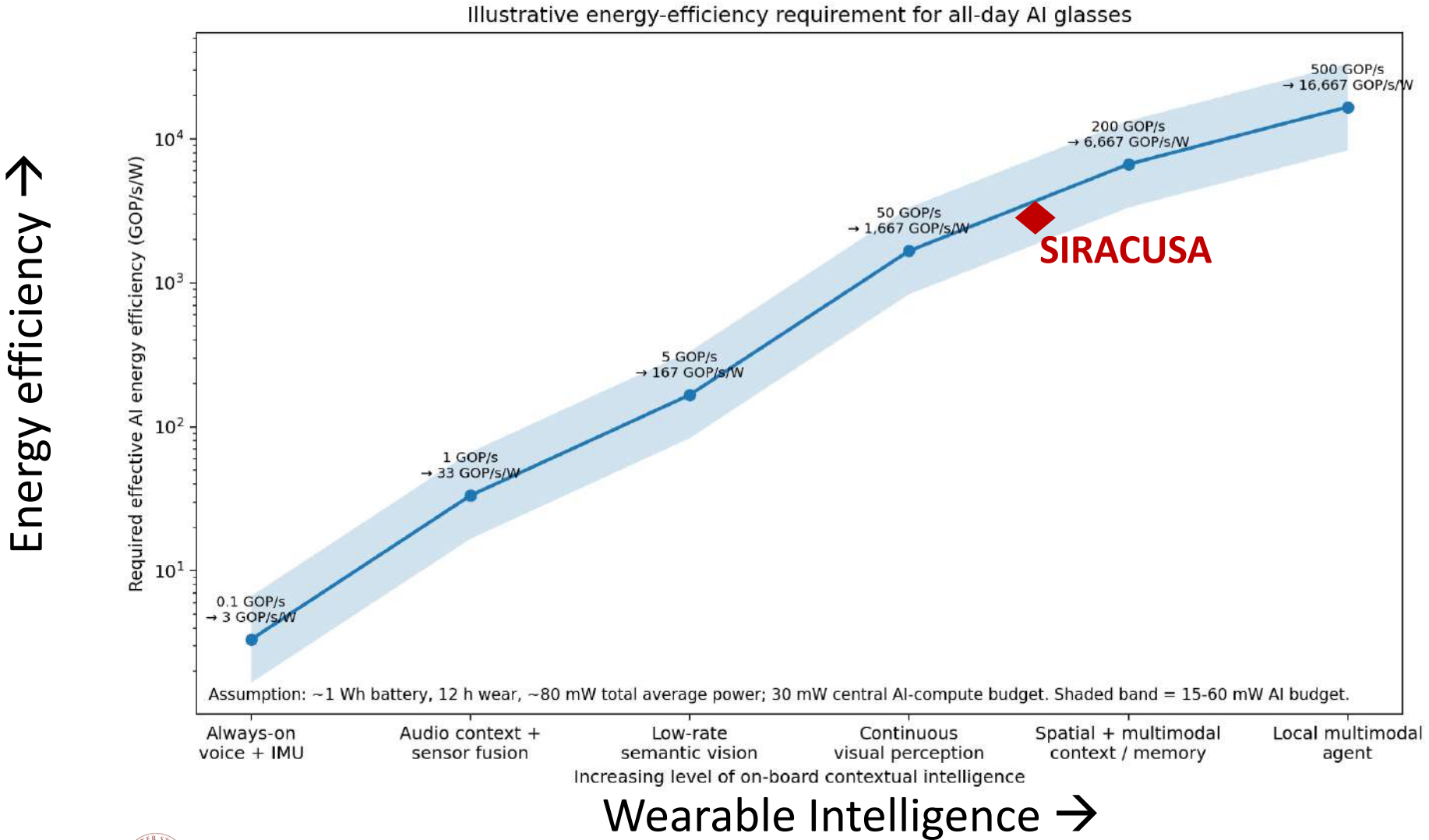
Multicast/reductions/compression

Compute near/in mem & NoC



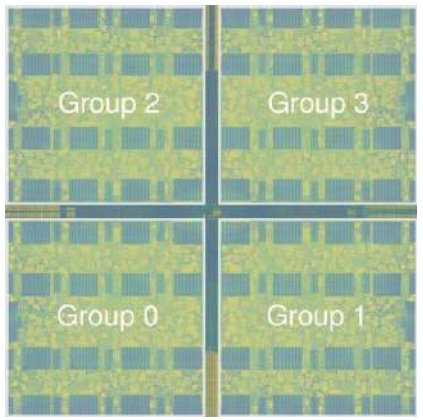
Efficiency Requirements for Wearable Intelligence

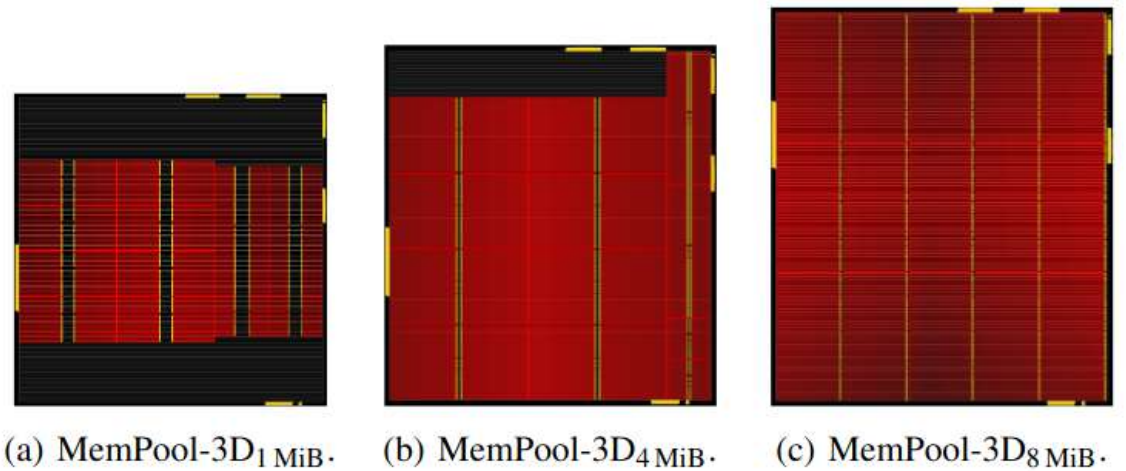
Where are we now?



3D for Scale Up!

- **Memory-on-Logic (MoL)** implementation of the cluster
- With 1, 2, 4, 8 MiB of L1 on a second stacked die
- More memory on the same x-y area

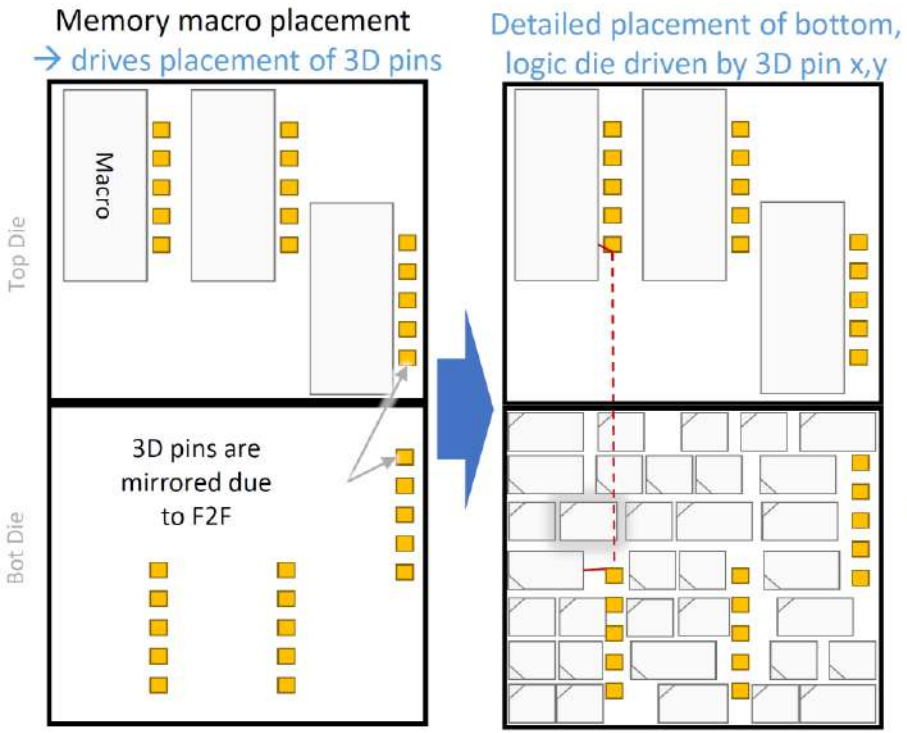


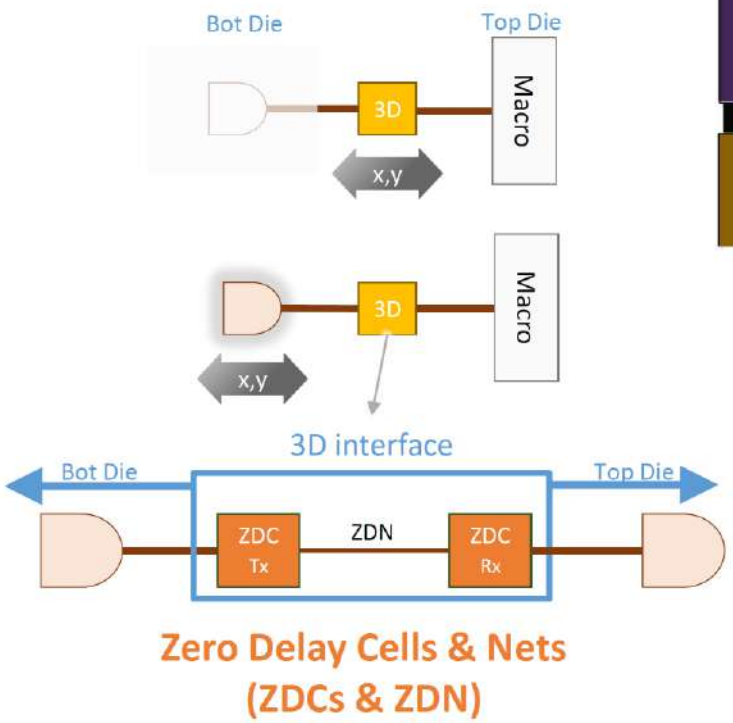


(a) MemPool-3D$_{1\,\mathrm{MiB}}$. (b) MemPool-3D$_{4\,\mathrm{MiB}}$. (c) MemPool-3D$_{8\,\mathrm{MiB}}$.

Design Tools Support is Essential

[IMEC, Synopsys DATE25]

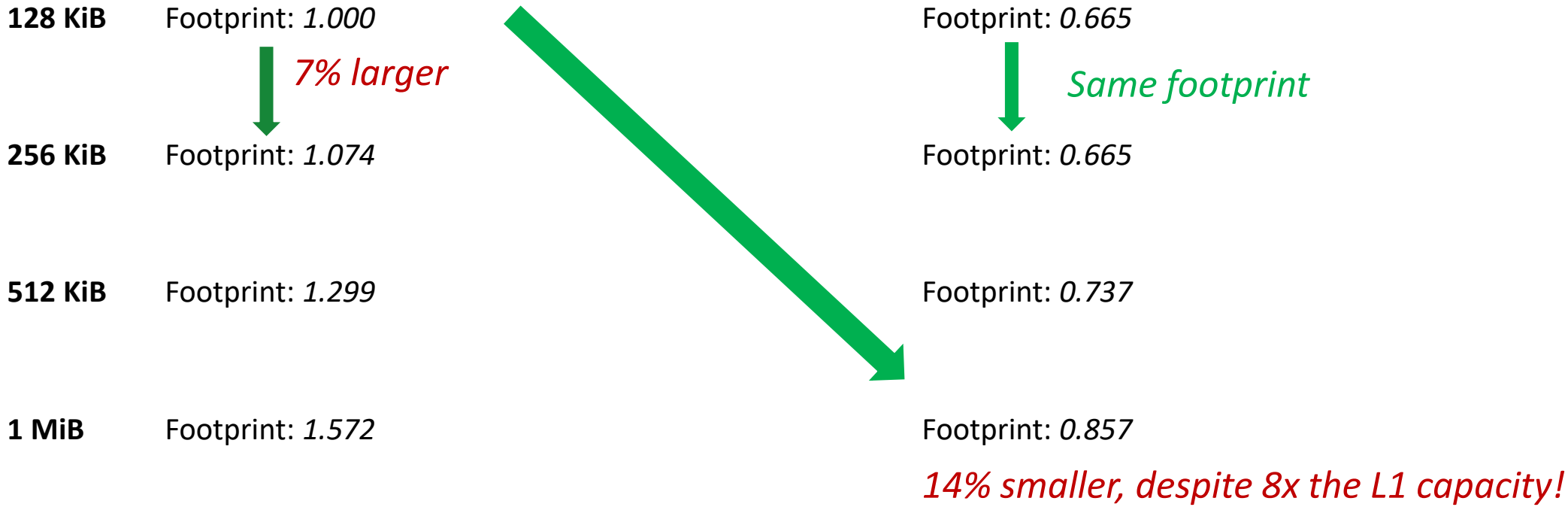




Accurate timing, power analysis

Full physical design (MoL F2F)

Implementation: 2D vs. 3D



Similar trends for **Wire length** and **#Buffers**

Can we Exploit the Extra L1?

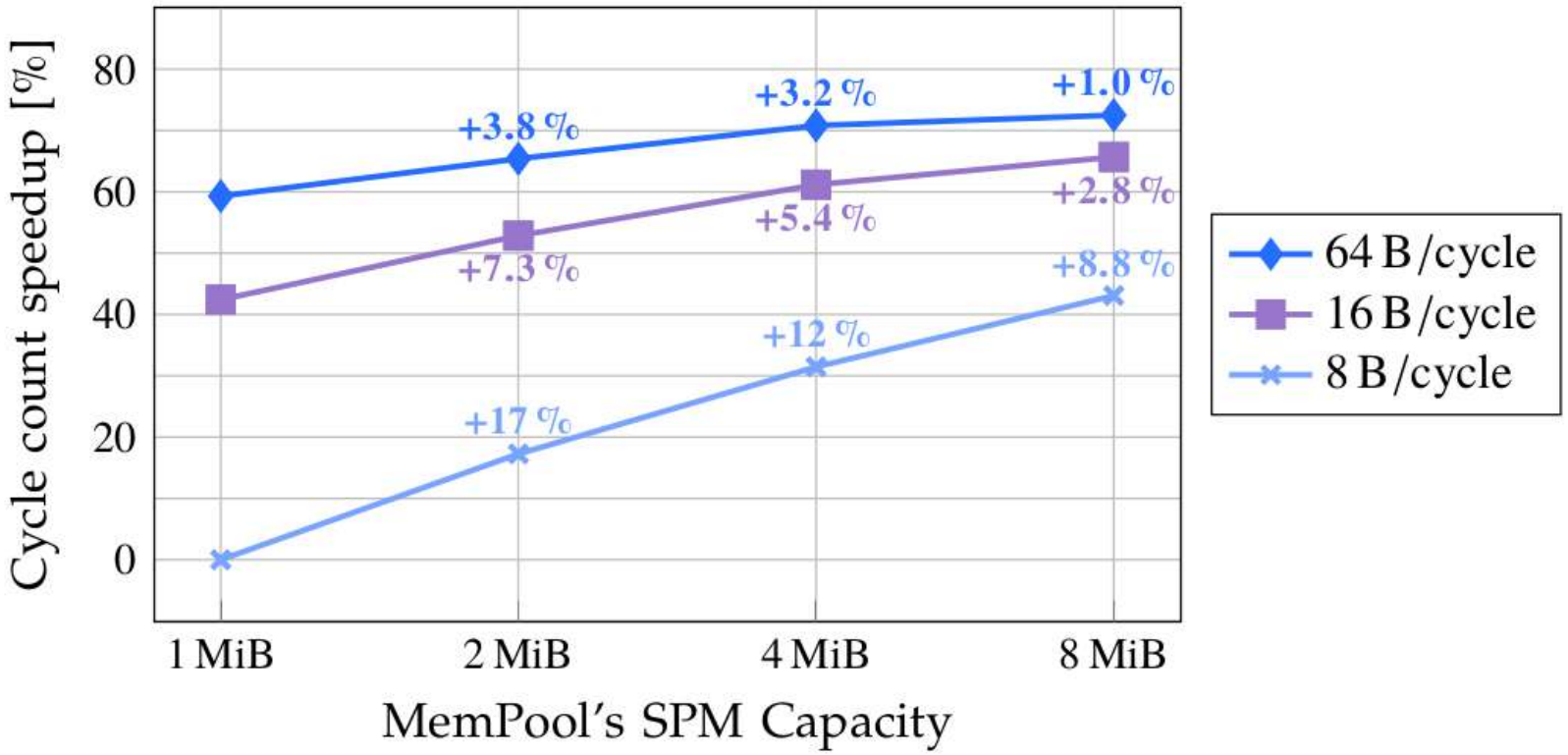
- **We can increase the L1 SPM capacity by 8, from 1 MiB to 8 MiB**
  - But does this have an impact in the cycle count performance results?

- **Benchmark: matrix multiplication of two large M × M matrices that do not fit in L1**
  - Cycle-accurate simulation with Verilator
  - M is chosen to enable optimal tiling for all configurations
  - *Larger L1 means larger tiling sizes and fewer loading phases*

- **Do the fewer loading phases matter to our cycle-count results?**
  - The answer depends on the bandwidth on the main memory

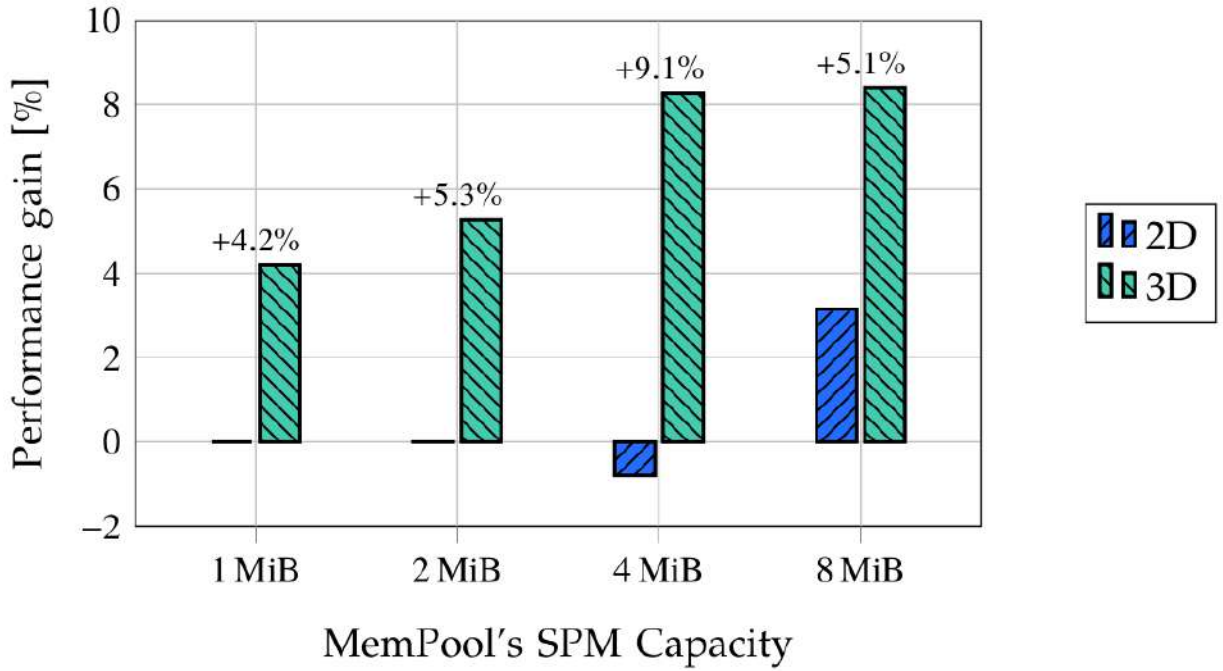
Matrix Multiplication: Cycle-Count

- **Low L2 Bandwidth**
  - Speedup of 42%
  - Memory-bound
- **High L2 Bandwidth**
  - Speedup of 8%
  - Compute-bound
- **Avg. L2 Bandwidth**
  - Speedup of 16%



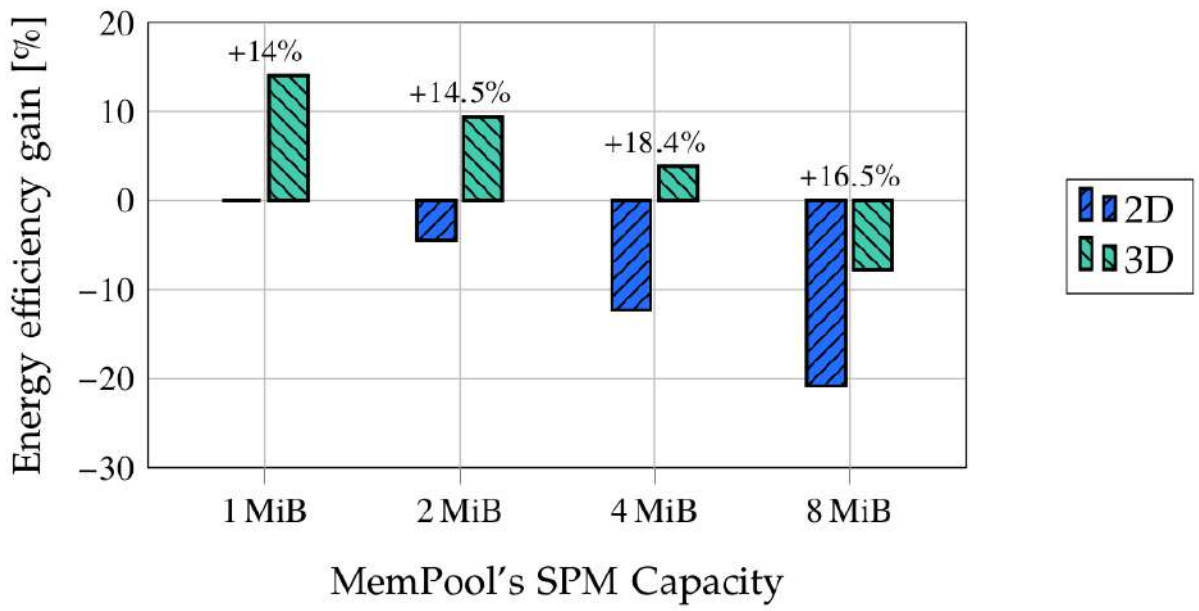
Performance

- **3D consistently outperforms 2D**
  - *Smaller footprint leading to higher frequency*
- **Larger L1 capacity → increased perf**
  - But only with 3D



Energy Efficiency

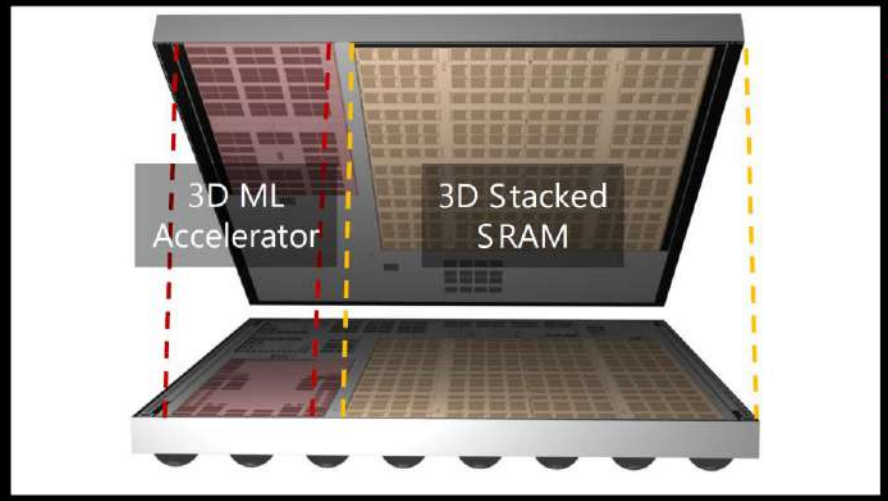
- **3D consistently outperforms 2D**
  - Smaller footprint leading to fewer buffers, shorter wire length, and smaller power consumption

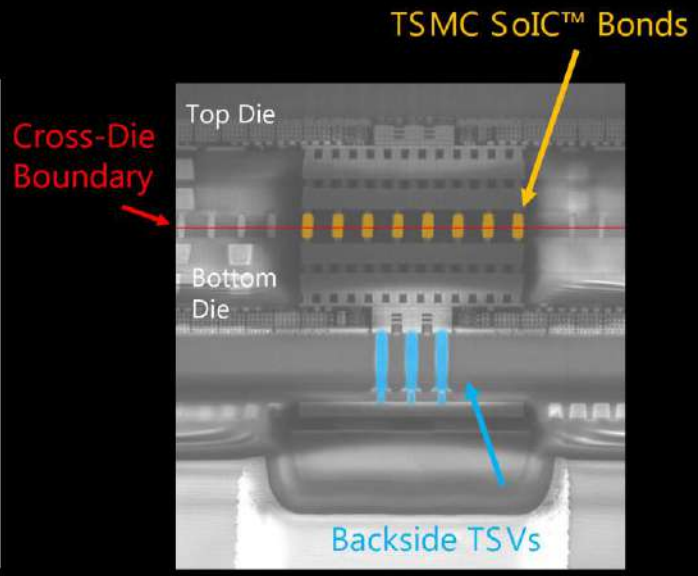


MoL is a significant scale-up booster!

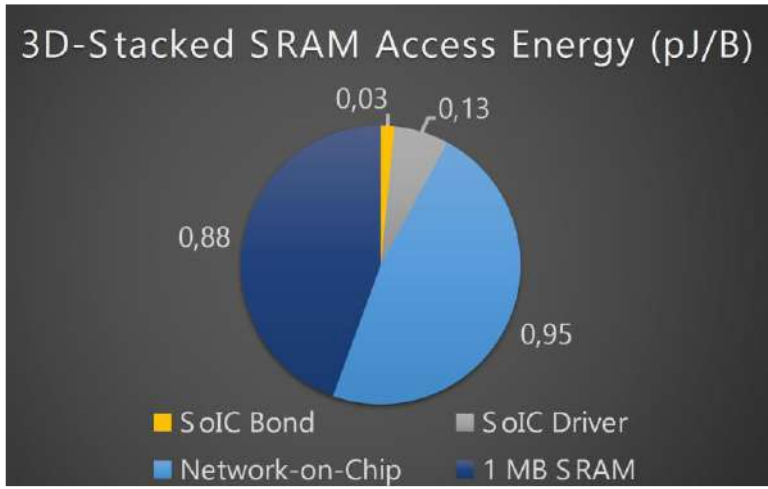
This is where industry is going too!

3D Integration Enables a Path to AR Silicon





- 3D stacking for addressing the combined Power-Performance-Area challenges of AR SoCs within the same 2D area footprint
- Implemented a prototype 3D stacked AR SoC: Two 7nm dies, wafer-on-wafer face-to-face stacked at < 2um bonding pitch using TSMC SoIC™ bonding technology
- Integrates 3D ML Accelerator, 3D Stacked SRAM, CPU for realistic model deployment



Meta – ISSCC24

Towards sub-mW always-on interaction

Event-driven perception through visual wake-up

Until now

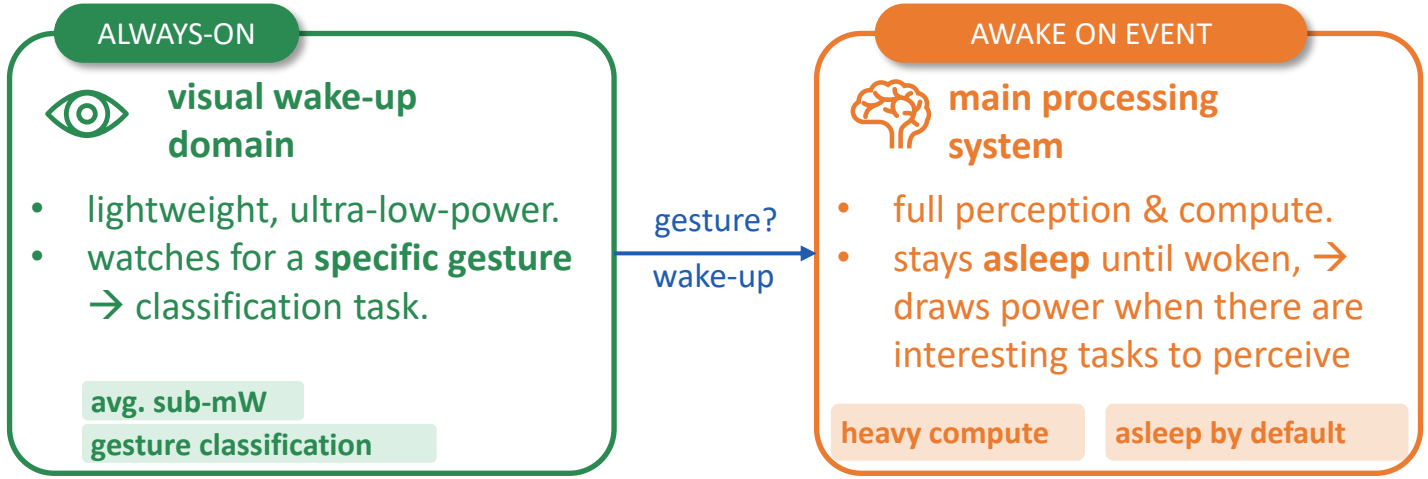
- a few **100 mW peak** inference power

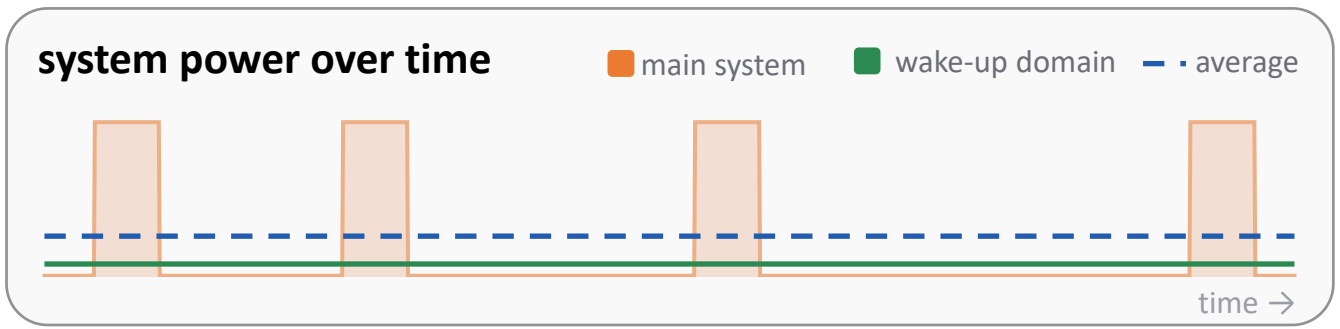
Is always-on rich perception necessary? **Often not** — e.g. sitting at a desk and working, rich perception is not needed.

Requirement

- low **10s mW average** power or lower

event-driven visual wake-up

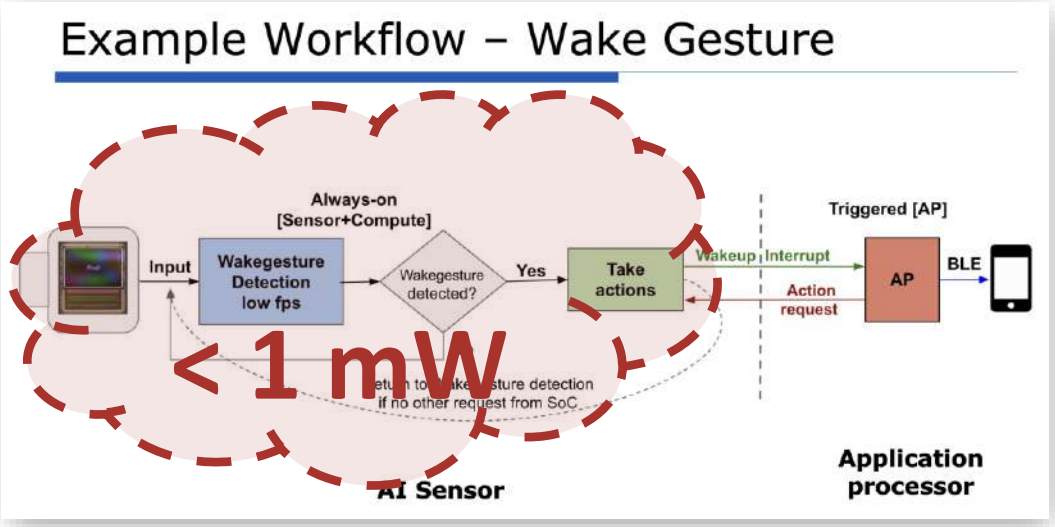




Wakelet: hierarchical sensing

- No fixed duty cycle
- Responsiveness
- Compute capabilities
- No waste of energy

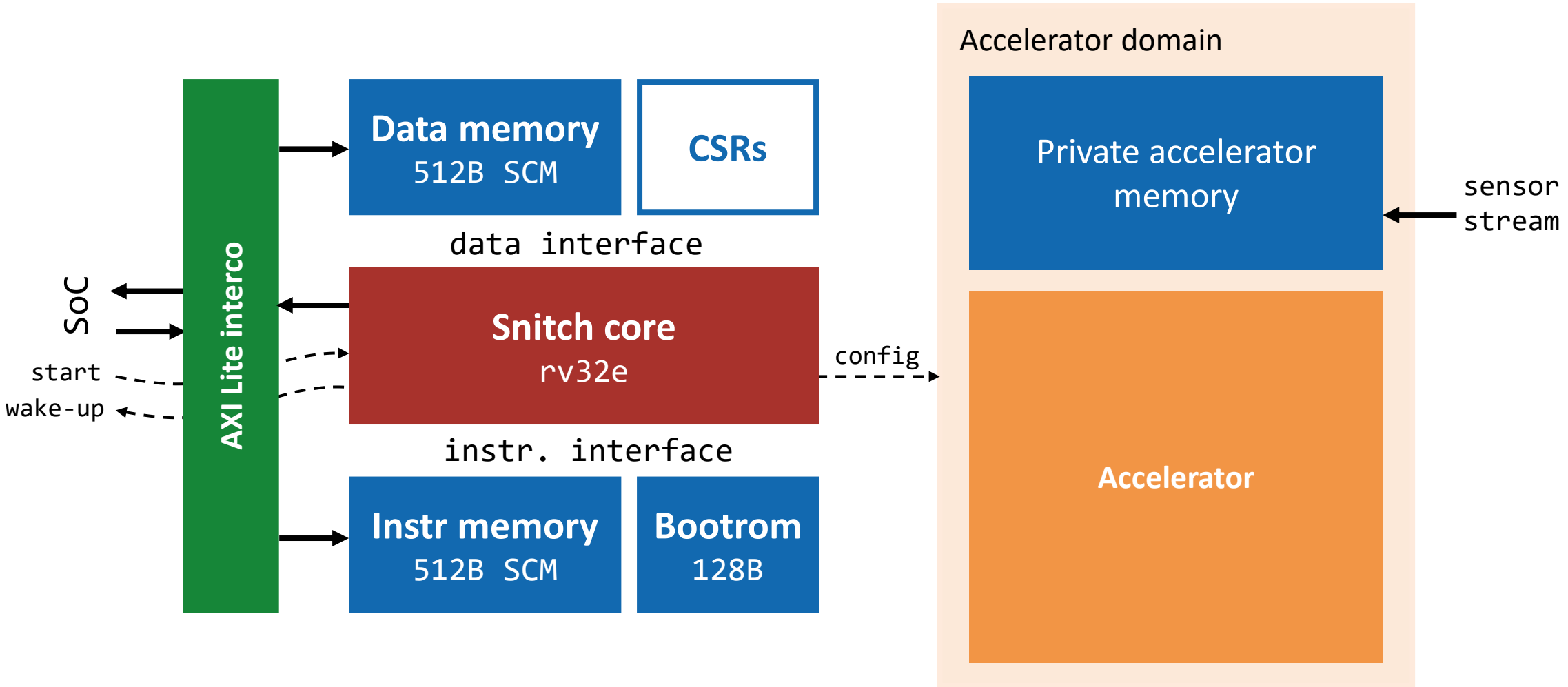
- **Wakelet**
  - Wrapper for always-on accelerators
  - Enhance accelerators with **programmability**
  - Negligible power consumption



Source: Meta, *ISSCC 2025* (Forum 5)

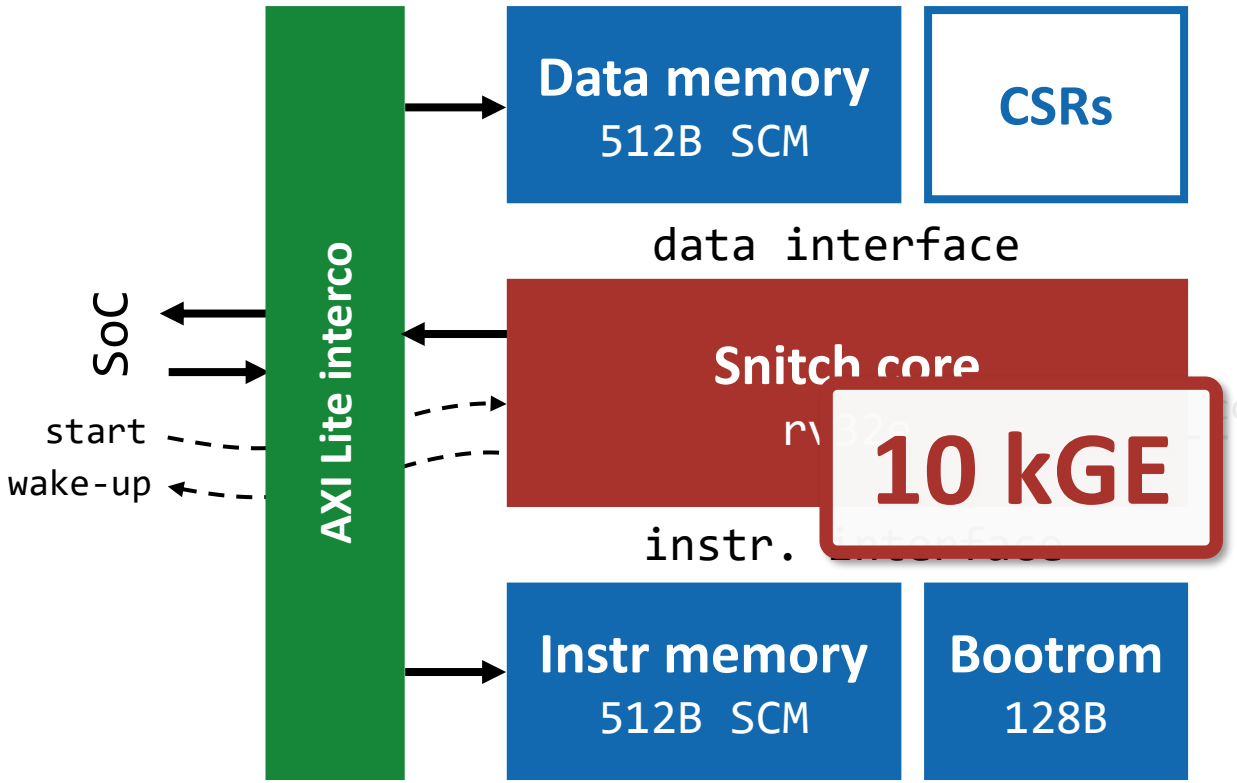
Sony: **790 µW** @ 1 fps high-res

Wakelet architecture



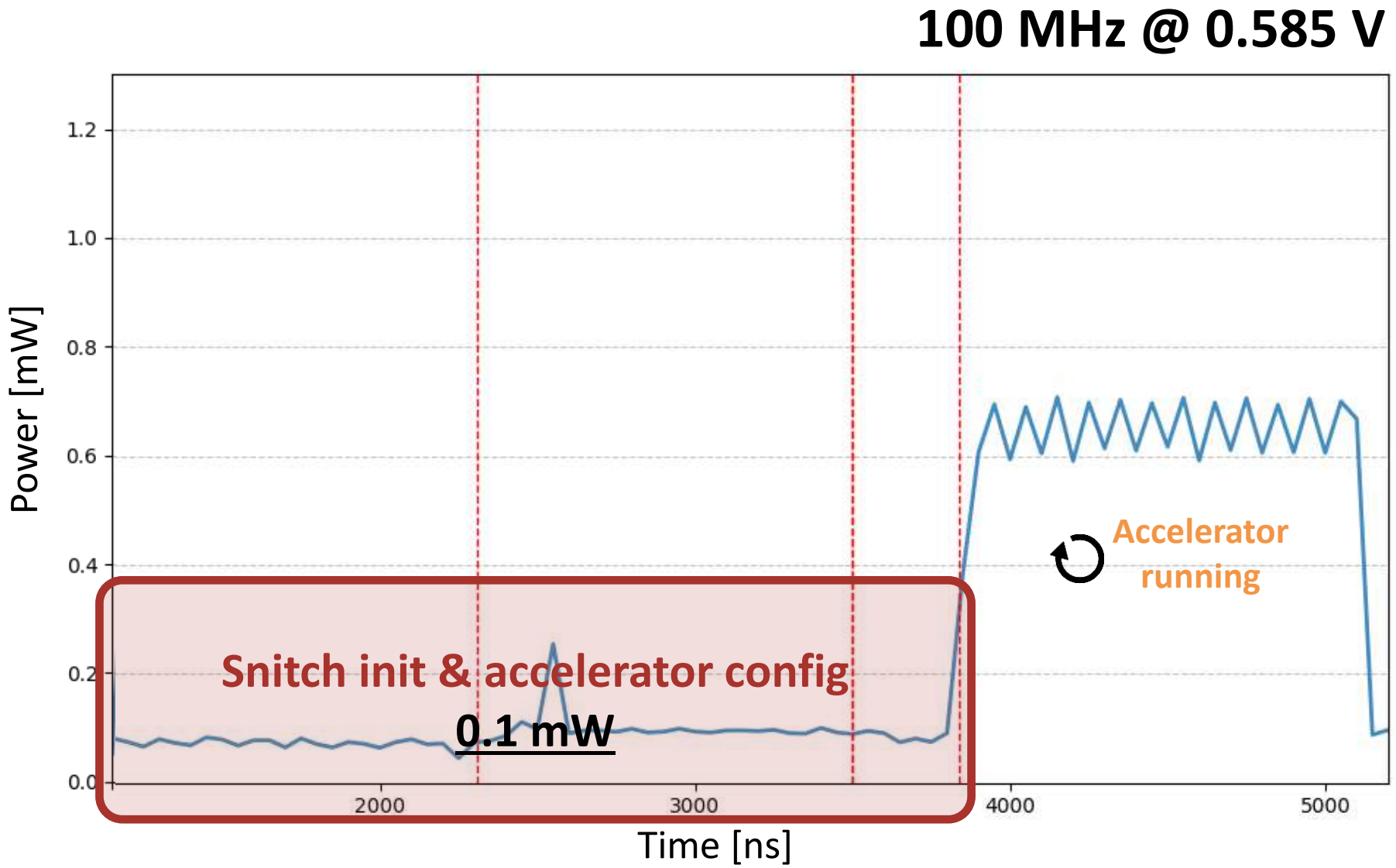
How big is Wakelet?

TSMC 7nm: 100 MHz/worst-case corner/only standard V_{TH}



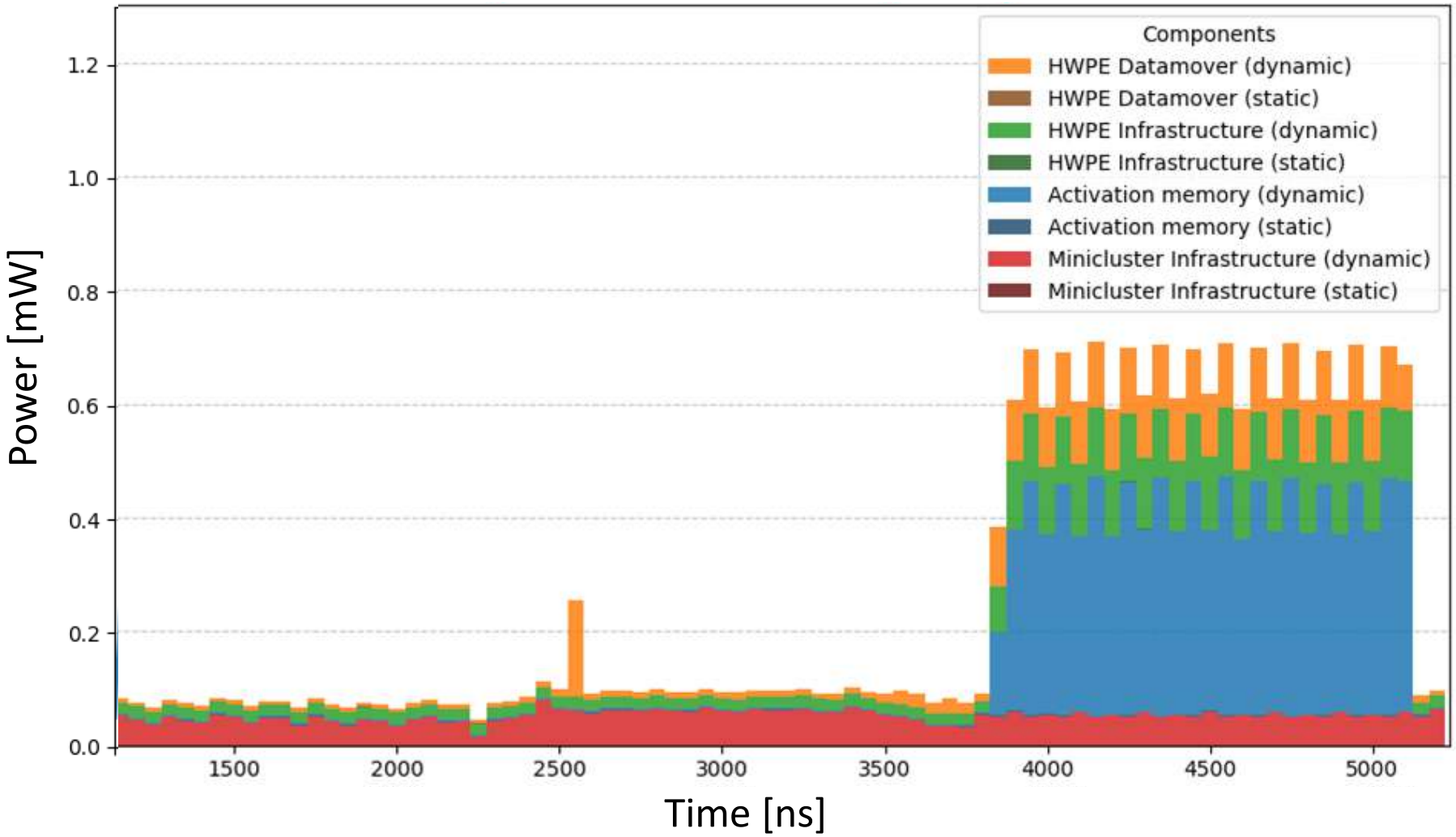
Only 60 kGE!

So, is it low-power for real?



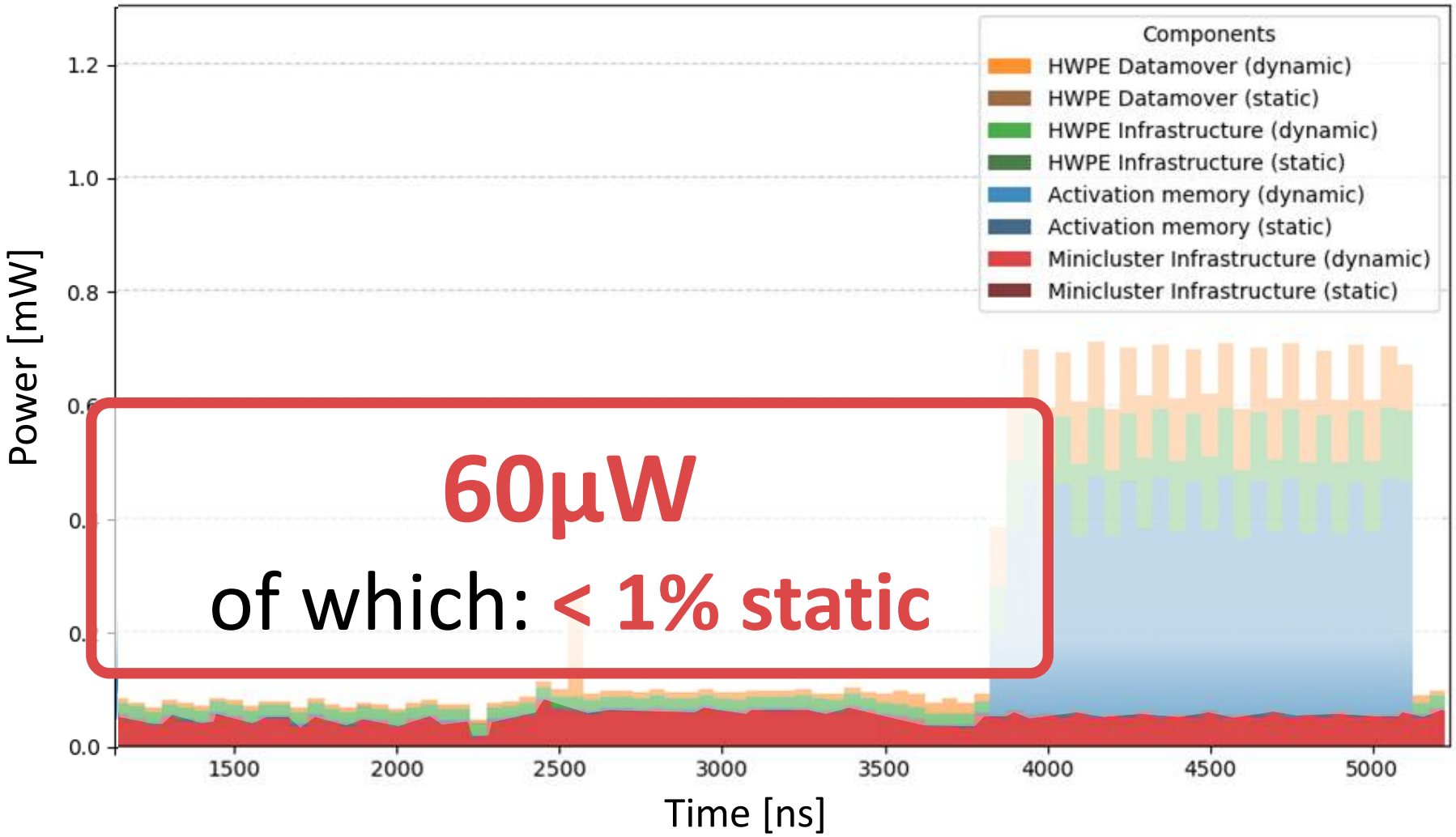
So, is it low-power for real?

100 MHz @ 0.585 V



So, is it low-power for real?

100 MHz @ 0.585 V

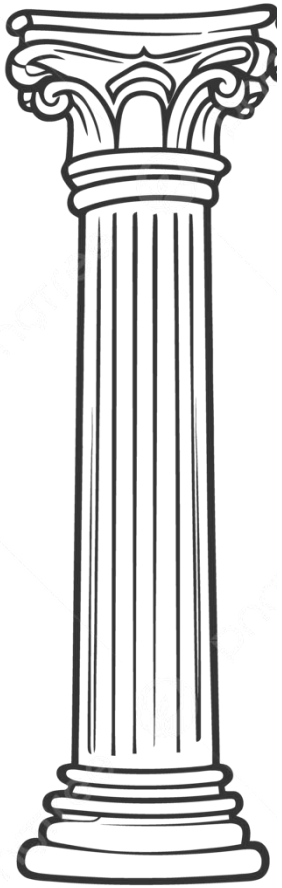


60µW

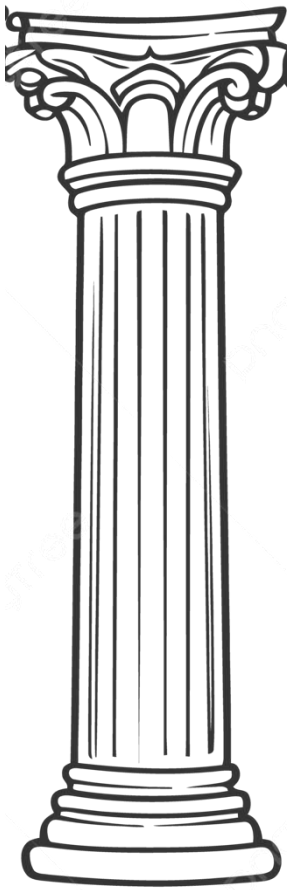
of which: **< 1% static**

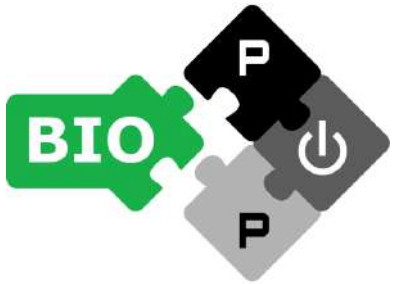
Not only Sound and Images

BIOPOTENTIALS

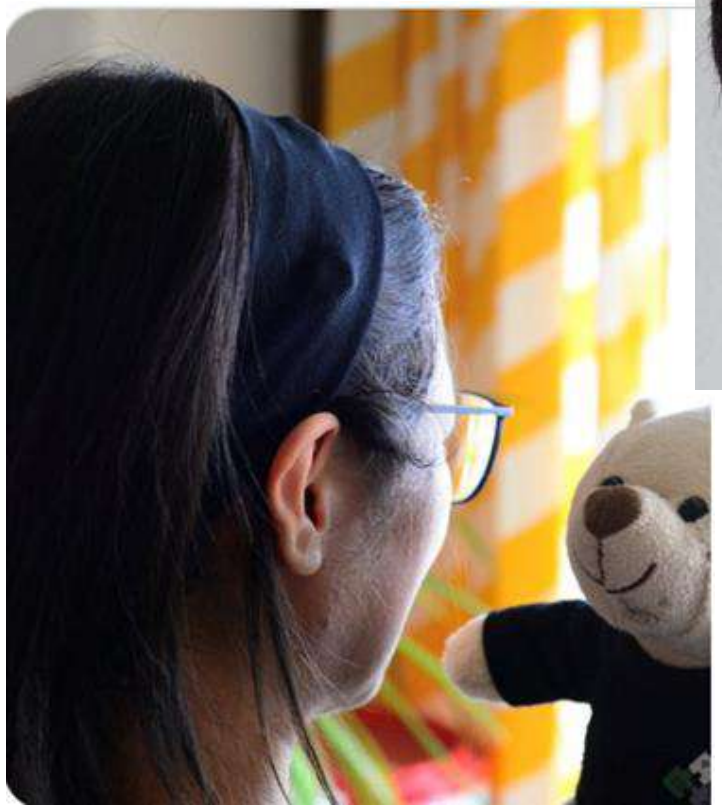


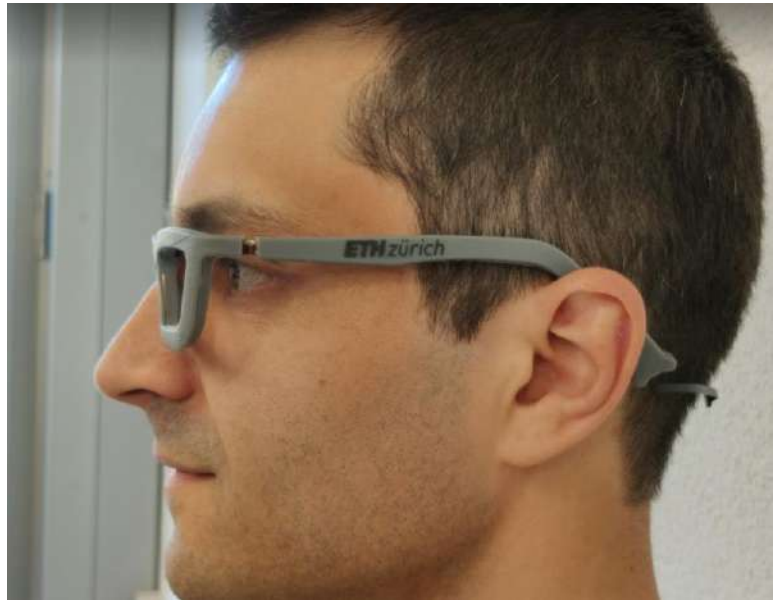
ULTRASOUND



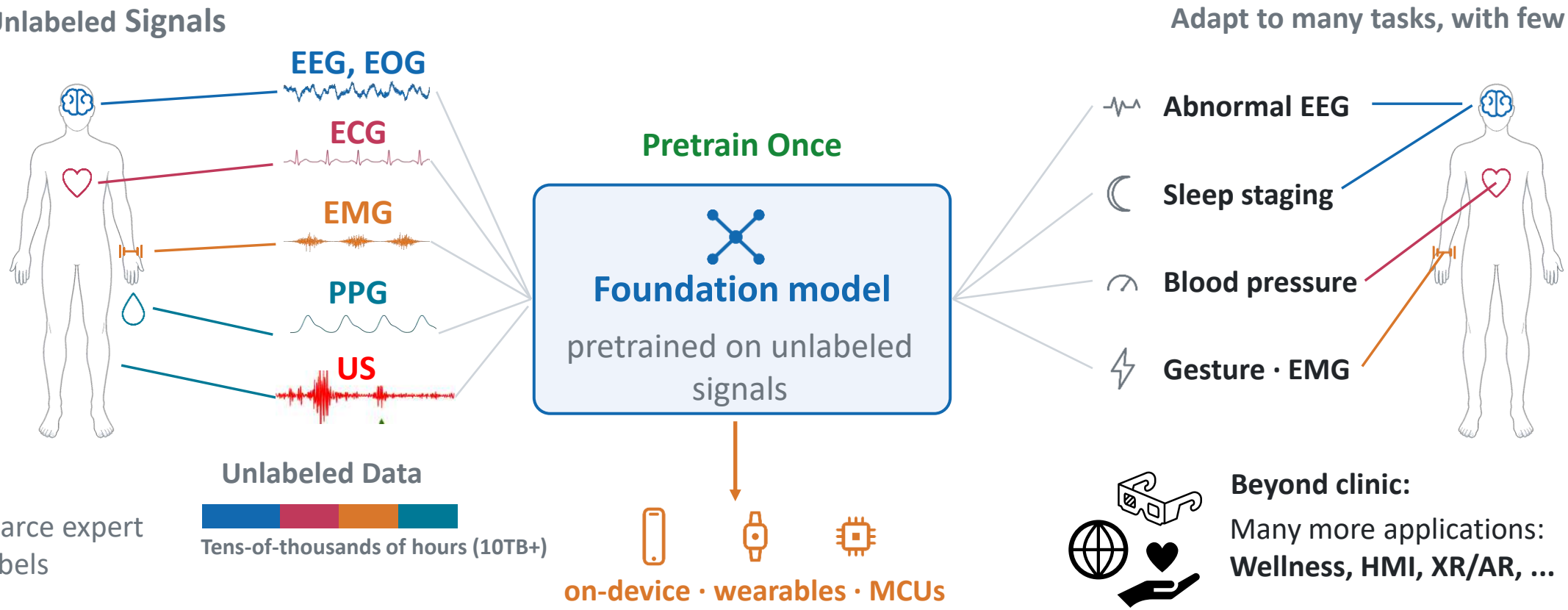


Not Only Smartglasses





Our dream: Foundation Model(s) for Biosignals & US



Pretrain one foundation model on unlabeled signals, then adapt it with few labels

The Future is AI-Boosted Extreme Co-design

Flexibility

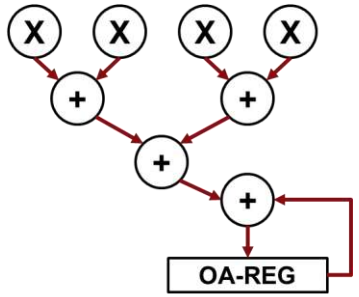
Energy-Efficient RV Core → **20pJ (8bit)**

ISA-based 10-20x →**1pJ (4bit)**

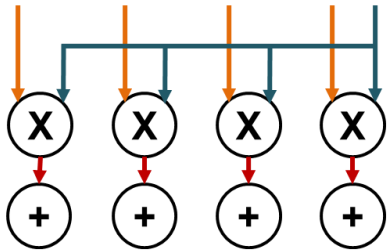
Configurable DP 10-20x → **100fJ (4bit)**

Highly specialized DP 100x →**1fJ (ternary)**

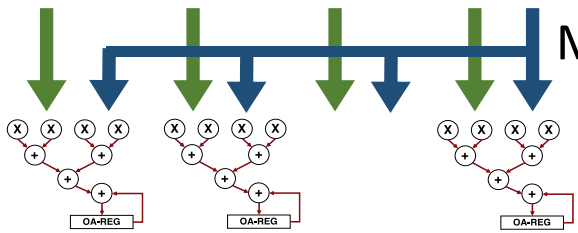
AI- boosted EDA



Inner Product



Outer Product



Mixed

Precision tuning – OP/Mem tuning - deep arithmetic optimization – 3D placement...

Co-Specialize SW, HW, EDA & Technology is the frontier & Open Platforms are Key!

Thank You!

youtube.com/pulp_platform

pulp-platform.org

@pulp_platform